Psychologie
Faculteit der Sociale Wetenschappen

Tracking Behavioral and Physiological Dynamics in Perceptual Decisions during Sustained Attention

Lucija Sorić

Research Master Thesis *Cognitive Neuroscience*

Date: 15/05/2026

Student number: s4507053

Supervisor: Dr. Anne Urai

Second reader: Prof. Sander Nieuwenhuis

Word count: 12 223

Abstract

We continuously rely on our ability to make decisions about incoming sensory stimuli. However, these decisions are highly variable, showing differences in accuracy and response time even when based on identical information. Although this variability is often attributed to random noise, recent work suggests it may reflect systematic processes, such as fluctuations in engagement states. Pupil diameter has been implicated as an indicator of these fluctuations. Previous research has reported both linear and quadratic relationships between baseline pupil diameter and behavioral engagement, depending on task structure and analytical steps. The present study used a threefold approach to reproduce, replicate, and extend these findings by examining perceptual decision-making and its psychophysiological correlates during sustained attention in a trial-based task without breaks. Seventy-two participants ($N = 72$) completed a 32-minute (min/max: 26.22 - 41.71 minutes) self-paced perceptual decision-making task in which they indicated which of two displayed stimuli had higher contrast. Participants were assigned to either a full-instructions condition ($n = 35$) or a minimal-instructions condition ($n = 37$). Behavioral and pupillary measures were analyzed as a function of time-on-task using linear and quadratic models. Across all participants, performance improved over time while pupil diameter decreased. No relationship between pupil diameter and performance was observed in the full-instructions condition. In contrast, the minimal-instructions condition showed a reliable inverted-U relationship between pupil diameter and response time metrics, which remained significant after accounting for time-on-task, and across time scales ($p \leq .001$). These findings do not support a consistent linear relationship between arousal and performance. Instead, they indicate that performance might be optimized at intermediate levels of arousal, and that this relationship is not uniform. This suggests that arousal–performance dynamics in perceptual decision-making could be context-dependent and may differ with increased task demands.

Keywords: perceptual decision-making, sustained attention, pupil diameter, replication, time-on-task

Layman's Abstract

We constantly rely on our senses to make decisions - recognising our name being called across a noisy room (sound), noticing that the milk smells off (smell), or spotting the emptiest train compartment at a glance (vision). Even in simple, familiar situations like these, we sometimes make mistakes or slow down and the reasons for this are not fully understood.

One explanation is that our levels of engagement naturally shift over time, affecting how well we perform. Pupil size offers a convenient window into these fluctuations - larger pupils reflect higher alertness and engagement.

Previously, it was found that performance is best at moderate levels of engagement, with both too little and too much alertness leading to mistakes. This relationship, first characterized in 1908, is widely known as the Yerkes-Dodson law. A recent study found evidence of it during a sustained attention task, and this thesis tested whether the same evidence can be found in a new context.

Seventy-two participants completed a 32-minute task in which they judged which of two images on a screen appeared brighter. Thirty-five participants were given clear instructions from the start, while the others had to figure out the rules on their own. Their choices and pupil sizes were recorded throughout.

Participants that knew the rules did not make worse decisions over time. In fact, their decisions improved, although their pupils got smaller. The same pattern was found for participants who had to figure out the rules by themselves. It was also noticed that their pupils were either very small or very large when their decisions were worse.

These findings support the idea that performance is best at moderate levels of engagement, particularly under challenging conditions. More broadly, they highlight that our decisions are not driven by any single factor, such as time, task demands, or alertness. Instead, they are a product of an interplay of a wide range of factors. This means that improving human decision-making might not be so simple, and future research needs to combine all of these factors to anticipate dips in performance.

Tracking Behavioral and Physiological Dynamics in Sustained Perceptual Decisions

Imagine standing in front of an apple stand at a bustling farmers' market. After some wait, it is finally your turn. In front of you are hundreds of apples, each slightly different in color, size, and ripeness. As you scan each apple on the stand, you decide whether to put it in your bag or leave it behind. Each decision then turns into a classification issue, requiring you to categorize every apple as "good" or "bad" based on certain visual features. This situation is an example of perceptual decision-making, a process that requires us to classify sensory information from the environment in order to select subsequent actions (Gold & Shadlen, 2007; Newsome et al., 1989; Sterzer, 2016; Summerfield & Blangero, 2017).

While such classification may appear straightforward, errors can still occur. For instance, an apple judged as "good" at the market may later turn out to be unripe or spoiled. Perceptual decisions can vary, even when based on identical information or strong sensory evidence (Duffy et al., 2025; Urai, 2025). In low-stakes contexts, such as selecting fruit, the consequences of this variability are typically minor. At worst, you might have to go a day or two without an apple. However, in high-stakes situations such as driving, even small fluctuations in perceptual decisions can have serious consequences. For example, a slight misjudgement of another vehicle's distance may determine whether you brake in time or cause a collision.

Despite the potentially serious consequences, fluctuations in perceptual decisions are common (Beck et al., 2012; Perquin et al., 2024). These fluctuations could be attributed to multiple sources, including external noise - that is, variability in the sensory environment itself (Beck et al., 2012; Ratcliff et al., 2018). For example, bright sunlight at a farmers' market may make it difficult to distinguish apples that are light in color due to illumination from those that are light because they are not yet ripe. However, fluctuations in decision accuracy and timing are also consistently observed in controlled experimental settings, where external noise is minimized or constant (Awwad Shiekh Hasan et al., 2012; Beck et al., 2012; Drugowitsch et al., 2016; Duffy et al., 2025). For this reason, understanding variability in perceptual decision-making requires consideration of other potential sources and/or influencing factors. The following sections review classic and contemporary accounts of decision variability, discuss factors that influence it, and outline the rationale and objectives of the present study.

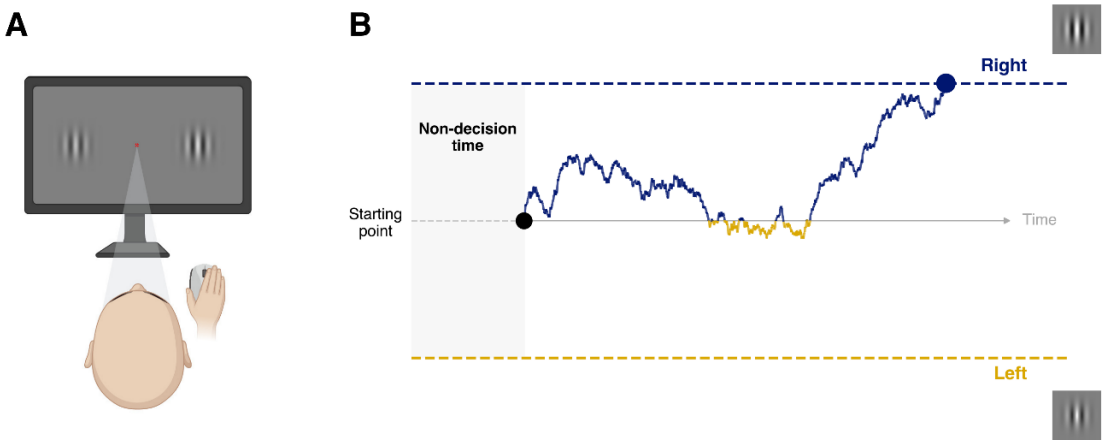
Sources of Variability

When making a decision, several elements are taken into account, including past experience (priors), current sensory information (evidence), and the consequences of the decision (value). Signal detection theory, one of the most influential frameworks in psychophysics, proposes that these components jointly contribute to a decision variable, which is then compared against a decision rule established based on task goals (e.g., choosing a ripe vs. spoiled apple; Gold & Shadlen, 2007; Green & Swets, 1966; Shadlen & Kiani, 2013). In signal detection theory, the sensory evidence that forms the decision variable is assumed to be noisy and this noise is considered the primary driver of

behavioral variability (i.e., differences in choices across identical trials). An extension of this is the sequential sampling framework, which does not rely on a single observation, but instead assumes that evidence is accumulated over time (Gold & Shadlen, 2007). Various models that incorporate both within-trial and between-trial variability have been developed within this framework (Duffy et al., 2025; Ratcliff, 1978; Ratcliff et al., 2016; Urai, 2025). The parameters of these models are particularly useful for modeling behavioral variability, as they account for two key measures of decision variability: choice accuracy and response time (Figure 1).

Figure 1

Example of a Sequential Sampling Model



Note. **A.** Illustration of the stimulus presentation for the task used in the present study. In each trial, participants were presented with two gratings on the screen and were required to indicate which one is of higher contrast. **B.** Schematic representation of the drift diffusion model for a single trial. The model assumes that binary decisions arise from the noisy accumulation of evidence over time (drift rate), depicted by the trajectory beginning at the starting point and terminating upon reaching a decision threshold (here, the subject decided "Right"). The moment at which the threshold is crossed defines the response time, which comprises both the evidence accumulation period and a non-decision time component (Mulder et al., 2012). In the illustrated trial, the upper threshold corresponds to the correct choice. The code to generate the model schematic is available on GitHub.

However, these classic frameworks typically do not account for the across-trial temporal structure of decision-making, which can be defined as behavioral patterns correlated over time (Avitan & Stringer, 2022; Boehm et al., 2018; Urai, 2025). As a result, they do not capture systematic fluctuations in variability that may unfold over the course of a task. Recent work suggests that such fluctuations may reflect shifts in internal states or decision strategies (Bolkan et al., 2022; Gilden et al., 1995; Roy et al., 2021; Wagenmakers et al., 2004). Internal states are latent processes that influence the way the brain responds to sensory inputs and generates behavioral outputs, thereby shaping decision-making (Flavell et al., 2022). This conceptualization is consistent with broader accounts of brain state, which have been described as patterns of ongoing neural activity that emerge from and have consequences for physiology and behavior (Greene et al., 2023; Nienborg &

Cumming, 2009). Although the current literature lacks consensus regarding the number or defining features of such states, constructs such as engagement, emotion, motivation, arousal, and homeostatic needs (e.g., hunger and thirst) have been proposed as internal states (Ashwood et al., 2022; Flavell et al., 2022; McGinley et al., 2015). These brain states are regulated by neuromodulatory systems (Figure 2.A), such as the serotonergic, dopaminergic and locus coeruleus–norepinephrine (LC–NE) systems, that can act across cortical networks and over relatively long timescales (Flavell et al., 2013, 2022; Grove et al., 2022; Grujic et al., 2024; Tsuda et al., 2026). Through this broad influence, neuromodulators play a central role in controlling internal states and their impact on behavior (McCormick et al., 2020). Their activity can be non-invasively tracked through pupillometry (Larsen & Waters, 2018), with pupil diameter linked to the LC–NE (Grujic et al., 2024; Joshi & Gold, 2020), serotonergic (Cazettes et al., 2021), and dopaminergic systems (Grove et al., 2022).

Sustained Decisions

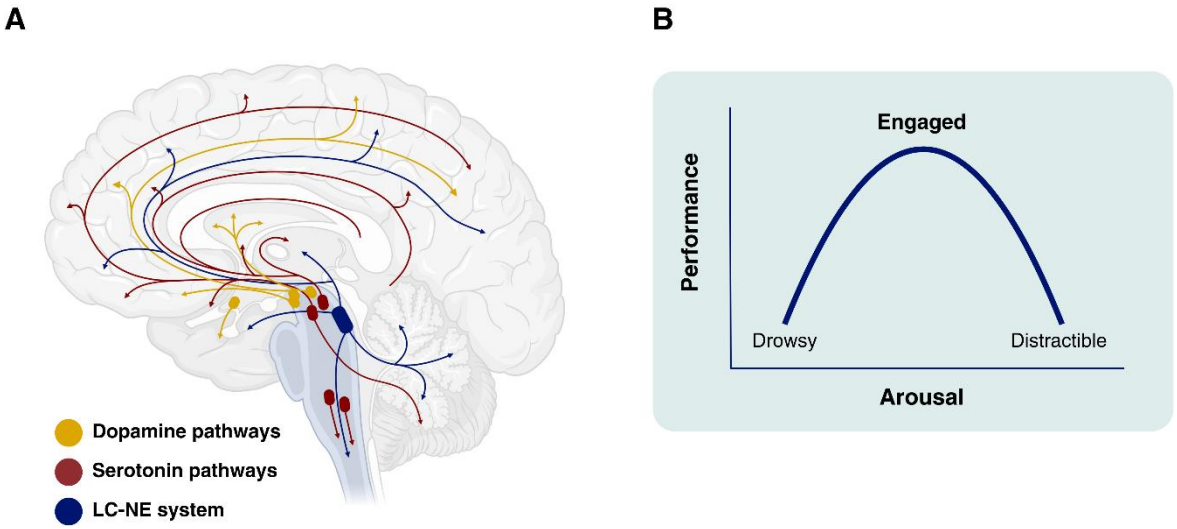
Across both real-world and laboratory settings, perceptual decision-making tasks often require individuals to exert attentional control over prolonged periods, assuming the ability to maintain engagement and performance throughout (Fortenbaugh et al., 2017; Langner & Eickhoff, 2013). This assumption, however, is challenged by evidence from both human and non-human populations. For example, mice have been shown to alternate between distinct behavioral states multiple times within a session, with lapses occurring predominantly during disengaged or biased states (Ashwood et al., 2022; Hulsey et al., 2024; Johnson et al., 2025). Similar fluctuations in performance have been observed in humans with increases in lapse rates, decreases in accuracy, and changes in response times (RTs) over the course of a task (van den Brink et al., 2016; Hopstaken et al., 2015a, 2015b). Together with task performance, pupil diameter has been identified as an external marker of internal states, capable of tracking fluctuations over time (van den Brink et al., 2016; Gilzenrat et al., 2010; Hulsey et al., 2024; Jepma & Nieuwenhuis, 2011; McCormick et al., 2020; Murphy et al., 2014). It is most commonly used as an index of arousal that reflects mental effort needed to allocate attentional resources to the task at hand (Alnaes et al., 2014; Bruya & Tang, 2018; Grujic et al., 2024; Kahneman & Beatty, 1966). Other constructs such as mental fatigue and vigilance have also been associated with fluctuations in pupil diameter (Hopstaken et al., 2015a, 2015b; Martin et al., 2022); however, these constructs are often used interchangeably in the literature and likely reflect similar underlying processes (Oken et al., 2006).

Pupillary dynamics have most commonly, but not exclusively, been linked to the neuromodulatory activity of the LC–NE system in both human and non-human mammals (Figure 1.A; Grujic et al., 2024; Joshi & Gold, 2020). Neural activity during LC stimulation has been related to changes in pupil diameter in rhesus monkeys (Joshi et al., 2016), and in mice (Reimer et al., 2016). In humans, converging evidence from neuroimaging and electrophysiology further supports this link: BOLD activity in the LC has been correlated with pupil dilation during tasks requiring increased mental effort, in both non-concurrent and concurrent pupillometry-fMRI designs (Alnaes et al., 2014;

Murphy et al., 2014). Pupil diameter has also been related to EEG markers of attentional state, such as the alpha-band amplitude and the P3 during decision-making, suggesting that momentary shifts in attentional engagement can be captured through pupillometry and electrophysiology (Hong et al., 2014; Montefusco-Siegmund et al., 2022; Murphy et al., 2011; Nieuwenhuis et al., 2005). Together, these findings suggest the potential of pupil diameter as a non-invasive, temporally sensitive index of LC–NE activity and arousal.

Figure 2

Some Neuromodulatory Systems that Control Internal States and the Yerkes-Dodson Law



Note. **A**. The locus coeruleus is a small nucleus (dark blue), located in the pons of the brainstem (light blue). It is the brain's primary source of the neurotransmitter norepinephrine, projecting widely across the cortex. The LC-NE system is crucial in regulating the arousal state (Aston-Jones & Cohen, 2005; Aston-Jones & Waterhouse, 2016; Breton-Provencher et al., 2021). Raphe nuclei (dark red) contain the majority of serotonergic cells and are located in the brainstem (light blue). Serotonergic projections are also widespread across the brain (Charnay & Leger, 2010). Dopamine is synthesized at various locations and projects across several pathways (yellow), such as the mesocortical, mesolimbic and nigrostriatal pathway (Klein et al., 2019). Created with BioRender. **B**. The Yerkes-Dodson law describing the relationship between performance and arousal, where arousal refers to an organism's level of responsivity to sensory stimuli (Pfaff et al., 2012).

Arousal across Time

Despite the potential of pupil diameter as an observable proxy of LC–NE activity, prior research reports conflicting findings regarding its relationship with performance during sustained attention. A commonly examined pupillary measure is the baseline pupil diameter, defined as pupil size prior to trial onset (e.g., during a fixation period), which reflects tonic, resting-state arousal (Aston-Jones & Cohen, 2005; Ayasse & Wingfield, 2020; Gilzenrat et al., 2010; Grujic et al., 2024). In mice, baseline pupil size has been shown to predict internal states and transitions between them, with both relatively small and relatively large baseline values associated with a higher probability of

disengagement (Hulsey et al., 2024; Johnson et al., 2025; Yerkes & Dodson, 1908). A similar pattern has been observed in humans where periods of disengagement have been linked to both larger (Murphy et al., 2014; Smallwood et al., 2011; Unsworth & Robison, 2016), and smaller baseline pupil diameter (Grandchamp et al., 2014; Hopstaken et al., 2015). Furthermore, sub-optimal task performance has been associated with both relatively small and relatively large baseline pupil sizes in some studies (Beerendonk et al., 2024; Murphy et al., 2011; van den Brink et al., 2016).

The literature thus reports contradictory findings regarding the relationship between baseline pupil diameter and task performance. Van den Brink et al. (2016) suggested that these inconsistencies may stem from the concealing effects of time-on-task, which could be masking a more complex relationship between baseline pupil diameter and performance. In the context of perceptual decision-making tasks, time-on-task effects refer to the linear changes in performance and physiological measures over time that occur during prolonged task engagement, often attributed to reductions in vigilance (McLaughlin et al., 2023; Warm et al., 2008). Consistent with this account, baseline pupil diameter has been shown to decrease across task blocks (Pielage et al., 2021; Unsworth & Robison, 2016). However, linear increases in pupil diameter during prolonged task engagement have also been reported (Murphy et al., 2011), and some studies have observed performance decline without corresponding changes in baseline pupil size (Beatty, 1982a, 1982b; Hopstaken et al., 2015a). These discrepancies may partly reflect differences in experimental design, which has been shown to moderate the relationship between time-on-task and the baseline diameter. For example, the inclusion of breaks between task blocks (McLaughlin et al., 2023) or reward manipulations (Hopstaken et al., 2015a, 2015b) has been found to alter tonic pupil activity, thereby influencing time-related changes.

Critically, when time-on-task was statistically controlled in studies without such experimental manipulations, the association between baseline pupil diameter and task performance was no longer linear but quadratic (van den Brink et al., 2016). This non-linear pattern is consistent with the Yerkes-Dodson law (Yerkes & Dodson, 1908), which describes performance as an inverted-U-shaped function of arousal (Hebb, 1955), with optimal performance at intermediate levels of arousal (Figure 2.B). It is also in line with the adaptive gain theory of LC–NE function which attributes this quadratic relationship to neuromodulatory dynamics within the system (Aston-Jones & Cohen, 2005). According to this framework, performance is poor at low tonic LC activity due to reduced alertness and at high tonic activity due to increased distractibility, but optimal at intermediate tonic levels when coupled with phasic LC responses to target stimuli (Aston-Jones et al., 1999; Aston-Jones & Cohen, 2005). Such relationship has been observed across species. In mice, performance in an auditory signal detection task was best at intermediate levels of arousal, as indexed by baseline pupil diameter (McGinley et al., 2015). Similarly, in a study examining internal states, mice showed the highest probability of being in an engaged state at moderate pupil diameters (Hulsey et al., 2024). In humans, the relationship between pupil-indexed arousal and task engagement has been observed in studies

examining the P3 event-related potential (Murphy et al., 2011), cortical spectral power (Podvalny et al., 2021) and behavioral performance (Beerendonk et al., 2024; van den Brink et al., 2016).

Current Study

Table 1

Comparison of the Paradigms Used by van den Brink et al. (2016) and the Present Study

	Gradual Continuous Performance van den Brink et al. (2016)	**Visual Decision-Making (Human IBL)** Enwereuzor et al. (2024)
Task structure	3 blocks x 8 minutes	1 block (~32.4 min average)
	5-minute break between blocks	No breaks
	600 trials	600 trials
	Continuous task	Separated trials with fixation
Stimuli	One stimulus at a time (detection)	Two stimuli simultaneously (discrimination)
	Stable stimulus contrast	Changing stimulus contrast
	Constant target location	Target location probability shifts every 20-100 trials
Conditions	Single condition	Two conditions: full vs. minimal instructions
Participants	N = 28	N = 72 (n_{full} = 35, n_{min} = 37)
	Right-handed only	Right- and left-handed
Hardware	EyeLink 1000	EyeLink 1000
	1000 Hz sampling rate	500 Hz sampling rate
Analysis	All trials included	Trials without contrast differences excluded
	Measures averaged per block	Single measure per participant
	40-second window (50 trials)	2.7-minute window (50 trials)
	MATLAB	Python

Note. The visual decision-making task used in the current study was designed to resemble the original mouse task as closely as possible. The main difference was that whereas mice indicated the spatial location of a single stimulus, humans identified which of two stimuli had higher contrast. The human task included two instruction conditions: a minimal condition, resembling the mouse experience (as mice always learn without instructions), and a full condition, more typical of human psychophysics tasks, allowing for comparison of behavioral differences between conditions.

A large body of evidence demonstrates that fluctuations in decision-making are common and may reflect systematic processes, such as changes in internal states (Ashwood et al., 2022; Bolkan et al., 2022; Duffy et al., 2025; Flavell et al., 2022; Gilden et al., 1995; Greene et al., 2023; McCormick et al., 2020; McGinley et al., 2015; Nienborg & Cumming, 2009; Roy et al., 2021; Urai, 2026; Wagenmakers et al., 2004). Given the ubiquity of such variability and its potential consequences, considerable research has focused on identifying the factors that drive fluctuations in decision-making and understanding the mechanisms through which they operate. While previous studies have suggested general principles, such as the inverted-U-shaped relationship between arousal and performance (Beerendonk et al., 2024; van den Brink et al., 2016), the robustness of this relationship across task contexts remains unclear (Nieuwenhuis, 2024). In particular, relatively little work has examined whether findings linking baseline pupil diameter to task performance can be replicated

across paradigms. To address this gap, I adopted a threefold approach to examine whether the findings by van den Brink et al. (2016) generalize to a different context (Bouter & Riet, 2021). First, I conducted a computational reproduction of the original findings, which refers to the assessment of the reliability of findings using the same methods and data (Parsons et al., 2022). The primary purpose of this step was to assess the validity of the analysis code developed for the subsequent replication. Second, I performed a conceptual replication by applying similar analytical procedures to a different experimental paradigm designed to study task engagement (Nosek et al., 2022). Finally, I extended this replication by exploring the relationship between behavior and pupil diameter using additional performance measures and experimental manipulations.

Whereas van den Brink et al. (2016) employed a continuous performance task consisting of 8-minute blocks separated by 5-minute breaks, the present study used a trial-based perceptual decision-making task without breaks, lasting on average 32 minutes. A full overview of the differences between the two designs is provided in Table 1. The task used in this study was developed as a human version of a standardized and reproducible perceptual-decision making task for mice (The International Brain Laboratory et al., 2021), which has been administered across a large number of mice and laboratories. This human adaptation enables direct cross-species comparisons of decision-making dynamics in humans and mice, and may provide a standardized measure of task engagement in humans. More broadly, this work is motivated by the growing emphasis on replication research in psychological science, which has been shown to promote positive structural and community changes within the field (Korbmacher et al., 2023; Nosek et al., 2022).

Building on the findings reported by van den Brink et al. (2016), I examined how time-on-task influences behavioral performance and baseline pupil diameter during sustained perceptual decision-making. Although van den Brink et al. (2016) did not explicitly state formal hypotheses, their expectations were outlined prior to the analyses. Based on these expectations, I tested the following hypotheses:

Hypothesis 1. Performance will decline over the course of the task. Specifically, participants will show an increased number of misses, and longer and more variable response times as time-on-task increases.

Hypothesis 2. Baseline pupil diameter will change over the course of the task, reflecting time-on-task effects.

Hypothesis 3. Baseline pupil diameter will initially show a linear relationship with performance in either direction. However, when controlling for time-on-task, this relationship will no longer be linear and a quadratic relationship will emerge.

In addition to testing these hypotheses, the design of the present study allows for exploratory analyses examining other factors that may influence decision-making during sustained attention. Specifically, the study employs a between-subjects design in which participants are randomly assigned to one of two conditions that differ in the level of provided instruction. In the full-

instructions condition, participants receive detailed instructions that clearly specify the task objective. In the minimal-instructions condition, participants are given essential information only (i.e., how to interact with the game, stating that the objective is to learn the rules by playing). This manipulation offers an opportunity to explore how differences in task understanding may influence behavior and physiology during sustained decision-making.

Methods

Reproduction

Design of the Original Study

In their study, van den Brink et al. (2016) used a modified version of the gradual continuous performance task. Participants (N = 28, M(age) = 20.9; SD = 2.5; min/max: 18–26; 6 male) were required to press the space bar upon detecting images of cities and to withhold responses when presented with images of mountains. Stimuli were presented continuously without breaks, transitioning between one another through morphing. Participants completed three blocks of 600 trials, with each block lasting approximately 8 minutes, separated by mandatory 5-minute rest periods. For a full description of the task, see van den Brink et al. (2016).

Data and Code Availability

A computational reproduction of the statistical results reported by van den Brink et al. (2016) was conducted in order to verify the validity of the developed code (Krafczyk et al., 2021). Although the original dataset has been made publicly available (van den Brink et al., 2017), the original source code was not. The analytical pipeline described in the original paper was therefore reconstructed in Python rather than the original MATLAB, allowing an assessment of the reproducibility of the reported findings. Reproducibility was assessed using the shared pre-processed dataset, which contained window-averaged values for each measure across three blocks per participant. These were subsequently block-averaged as reported in the original study. It was unclear from the pre-processed dataset whether the measures had been z-scored prior to sharing.

Replication on Novel Dataset

Participants

A total of 96 individuals participated in the study. The final sample size was determined following data quality screening of the recorded experimental sessions. Specifically, pilot sessions, sessions affected by technical errors, early terminations, sessions with unusually long durations, and duplicate participation were excluded from the analyses. A detailed overview of the data inclusion criteria is provided in Appendix A.

After applying these criteria, the final sample consisted of 72 participants (M(age) = 19.93; SD = 3.26; min/max: 18–35; 11 male, 1 non-binary). Most participants were right handed (N = 63). All participants provided informed consent prior to participation and reported being healthy adults with normal or corrected-to-normal vision and no current psychiatric, neurological, or psychological

disorders, thus meeting the inclusion criteria reported in the ethics application. (reference number 2024-03-12-A.E.Urai-V1-5358).

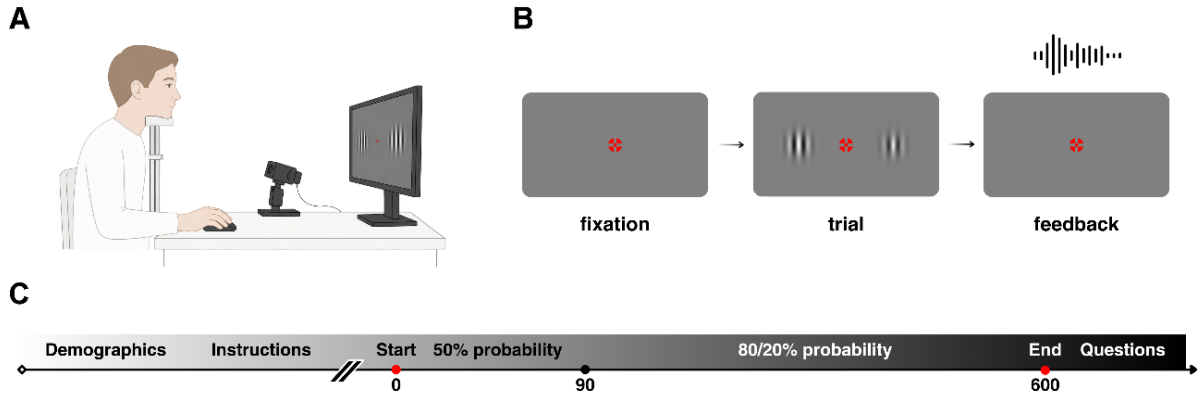
The study by van den Brink et al. (2016) did not report effect sizes. However, converting the reported t-statistic ($t(27) = 1.99$) to Cohen's d using Equation 7 from Lakens (2013) yields an effect size of $d_z = 0.38$, indicating a small-to-medium effect. A total sample size of 72 participants (35 and 37 per group) should therefore provide sufficient power to detect an effect of a similar magnitude.

Visual Decision-Making Task

The visual decision-making task used in the present study was developed by Enwereuzor et al. (2024) and administered using PsychoPy 2024.1.4 (Peirce et al., 2019). The task was adapted from a paradigm designed to study decision-making in mice (International Brain Laboratory et al., 2021). It was developed as part of a larger cross-species project comparing human behavioral, physiological, and neural data with archival mouse data. In the present study, the task difficulty was modified to enable valid cross-species comparison while avoiding ceiling performance and accounting for superior visual acuity in humans (Prusky & Douglas, 2004; Schnabel et al., 2021). Whereas mice were required to indicate the spatial location of a single stimulus (left or right), the human version required participants to identify which of two simultaneously presented stimuli was of higher contrast (Figure 3.A).

Figure 3

Overview of the Set-up, Trial Structure, and Task Structure



Note. **A**. Experimental set-up. Participants were seated at a standardized distance from the monitor and the eye-tracker, and were instructed to fixate their eyes on the fixation cross. Illustration generated with assistance of GPT Image 2.0. **B**. Example of a trial with the target stimulus on the left. **C.** Schematic overview of the experimental procedure.

Experimental conditions. Prior to starting the task, participants were randomly assigned to one of the two between-subjects conditions: a full-instructions condition ($n = 35$), and a minimal-instructions ($n = 37$) condition. Participants in the full-instructions condition received detailed written instructions, example stimuli, and completed five practice trials. Participants in the minimal-instructions condition received only the information necessary to perform the task (e.g., how to

respond). The instruction texts are provided in Appendix B. The two condition groups did not differ in gender distribution (χ^2 (2, N = 72) = 1.10, p = .576), nor the mean age (t(70) = 0.69, p = .492). The group-wise distribution of demographic variables can be found in Appendix C.

Stimuli. On each trial, participants were presented with two Gabor patches (gratings) displayed simultaneously on the screen, one to the left (-618, 0) and one to the right (618, 0) of a central fixation cross. The size of the gratings was 618 x 618 pixels. The gratings were of sinusoidal wave embedded in a Gaussian envelope, with a spatial phase randomly generated on each trial, so that the sinusoidal pattern appeared in a different position across trials. Spatial frequency was held constant at 0.0073 cycles per pixel. The target stimulus was defined as the grating with higher contrast. Target contrast levels ranged non-linearly from 0.50 to 0.70 (0.50, 0.52, 0.55, 0.60, 0.70), while the non-target stimulus always had a contrast of 0.50. A red fixation cross was displayed at the center of the screen (0, 0) throughout the experiment.

Trial structure. Each trial began with a fixation period lasting between 0.4 and 0.7 seconds. Stimulus onset was accompanied by a brief auditory cue (0.1s, 5000 Hz). Following stimulus presentation, participants indicated their choice by sliding a grating to the center of the screen using the cursor. A response was registered once the center of the selected grating reached the center of the screen, or the equivalent distance off the screen. Participants had up to 10 seconds to respond. If no response was made within this time window, the trial was classified as a timeout. Timeouts were followed by a 0.5s low-pitched tone (567 Hz). Correct responses (selection of the higher-contrast grating) were followed by a 0.2s high-pitched tone (2000 Hz). Incorrect responses triggered a 0.5s white-noise burst and timeouts triggered a 0.1s feedback sound (200 Hz). Incorrect responses and timeouts were followed by a 1s inter-trial penalty. Figure 3.B shows an illustration of the trial structure.

Task structure. The experimental task consisted of 600 trials. During the first 90 trials, the target stimulus appeared on the left or right with equal probability (50%). In the remaining 510 trials, the target stimulus appeared on the left or the right with either 20% or 80% probability, depending on a pre-generated sequence. This structure follows the design of the original task in mice (International Brain Laboratory et al., 2021). A schematic overview of the task structure can be found in Figure 3.C. The task was self-paced and lasted on average 32.40 minutes across the full sample (min/max: 26.22–41.71 minutes). Participants in the full-instructions condition completed the task slightly faster (M = 31.79 minutes, min/max: 26.22–40.29) than those in the minimal instructions-condition (M = 32.99, min/max: 26.30–41.71). This difference, however, was not statistically significant (t(70) = 1.23, p = .224).

Performance measures. The primary measure of performance was lapses of attention, operationalized as false alarms. In the present study, false alarms were defined as trials in which participants made a perceptual error by selecting the grating with lower contrast. In line with van den Brink et al. (2016), additional measures of task performance were included: (1) mean RT, (2) the

proportion of trials falling within the slowest quintile of RTs, and (3) the response time coefficient of variation (RTCV), calculated as the standard deviation of RT divided by the mean RT. Prior to computing these measures, a sliding-window approach was applied to each participant's behavioral data. Specifically, a window of 50 trials was moved across the dataset in steps of 15 trials, resulting in 37 overlapping windows. A schematic overview of the sliding-window approach can be found in Appendix D. For each window, the following metrics were calculated: (1) the proportion of false alarms, (2) the mean RT, (3) the proportion of trials within the slowest RT quintile (relative to each participant's RT distribution), and (4) RTCV, computed as the standard deviation of RT within the window divided by the participant's overall mean RT. This procedure yielded continuous time series of performance metrics, which were subsequently standardized (Z-scored).

Pupillometry

During the visual decision-making task, participants' pupil activity was recorded using an SR Research EyeLink 1000 eye-tracker. Monocular recordings of the right eye were acquired using a 35 mm camera lens in pupil-CR tracking mode at a sampling rate of 500 Hz. Participants' head position was stabilized using a chinrest without a forehead rest (Figure 3.A). Prior to the experiment, the eye-tracker was calibrated using a standard 9-point calibration procedure, followed by validation. Communication between PsychoPy and the eye-tracker was established using the EyeLink plugin for PsychoPy.

Pre-processing. Pupil data were pre-processed by Johnson et al. (2026) using MNE-Python. The raw data contained periods of missing signal automatically flagged by SR Research's blink detection algorithm. A secondary algorithm was then applied to identify additional moments of low signal based on detecting sharp changes in pupil size. All flagged periods were interpolated, and the continuous signal was subsequently bandpass filtered (0.01 - 10 Hz). Finally, a linear regression was used to remove pupil variance associated with blinks and saccades resulting in a continuous pupil time series. For the present study, the pre-processed signal was segmented using event markers sent to the EyeLink system during task acquisition. For each trial, I extracted the average of the pupil diameter recorded during the fixation phase. There were no trials in which a blink occurred throughout the entire fixation period, therefore no trials were excluded.

Pupillary measures. Prior to computing pupillary measures, a sliding-window approach identical to the one used for the behavioral data was applied to the trial-averaged pupil time series. Specifically, a window of 50 trials was moved across the data in steps of 15 trials, resulting in 37 overlapping windows. For each window, two measures were calculated: (1) the mean baseline pupil diameter, and (2) the mean temporal derivative of the baseline pupil diameter, reflecting the rate of change in pupil size (i.e., pupil constriction or dilation) within the window. Negative values of this derivative indicate that the pupil was, on average, constricting during the window, whereas positive values indicate dilation. The resulting pupillary time series were continuous, matched the length of the behavioral data, and were subsequently Z-scored.

Procedure

Upon arrival at the laboratory, participants received written and verbal information about the study, including an overview of the procedure and experimental design, eligibility criteria, compensation, confidentiality provisions, and contact details for further inquiries. All participants provided written informed consent prior to participation and were reminded of the voluntary nature of their participation. Participants were then seated in a dimly lit room, and were instructed to rest their chin on a chinrest. The video camera was activated and the EyeLink 1000 Plus eye-tracking system was calibrated and validated. Before beginning the experimental task, participants reported their age, gender, and handedness. They were subsequently randomly assigned to one of two experimental conditions. Depending on the assigned condition, participants received either full or minimal task instructions. Participants in the full-instruction condition additionally completed five practice trials to familiarize themselves with the task. Following the practice trials (if applicable), participants completed the main experimental task, which lasted on average 32.4 minutes (26.22–41.71 minutes) and consisted of 600 trials. In each trial, participants chose between two visual gratings presented on the screen that varied in contrast by sliding one to the middle of the screen using a mouse. After completing the task, participants answered a brief set of questions regarding their experience and were debriefed. Compensation was provided at the end of the session in the form of either SONA credits (2 credits) or monetary payment (€7.50). If participation exceeded the scheduled duration of 60 minutes, additional compensation was provided in accordance with CEP guidelines. The study protocol, including the procedure described above, was approved by the Psychology Research Ethics Committee of the Institute of Psychology on 12 March 2024.

Statistical Analyses

Behavioral data preprocessing was conducted in R 4.5.1, while pupil preprocessing and all subsequent analyses were performed in Python 3.13 (versions 3.13.5 and 3.13.9, depending on library requirements), using libraries including NumPy, Pandas, and Matplotlib, among others. A complete list of dependencies, along with all analysis code, is available on GitHub.

The analysis pipeline follows the approach described by van den Brink et al. (2016), whereby analyses were conducted at the participant level with trial-level measures aggregated per participant for group-level analyses. The first hypothesis predicted performance decrements over time, specifically an increased number of lapses, slower and more variable RTs. To test this, one-tailed t-tests were used to assess whether the slopes of fitted regression lines showed significant increases over the course of the task. The second hypothesis predicted systematic changes in pupil diameter over time and was evaluated using two-tailed t-tests. The third hypothesis predicted a quadratic relationship between performance metrics and pupillary measures after controlling for time on task. This was examined by comparing the distribution of regression coefficients against zero, using one-tailed t-tests for baseline pupil diameter and two-tailed t-tests for the pupil diameter derivative.

For the reproduction and replication analyses, one-sample t-tests were used unless otherwise specified, consistent with van den Brink et al. (2016). Exploratory analyses comparing means across the two experimental conditions employed Welch's t-test, given the skewed distributions of RTs and unequal group sizes. The exploratory analyses proceed in two directions: one applies the same pipeline as van den Brink et al. (2016) to the minimal-instructions condition, while the other directly compares performance measures between the two conditions.

Given the multiple comparisons conducted within each hypothesis of the replication analyses, p-values were corrected for false discovery rate using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995), applied separately within each hypothesis family. This correction was also applied to the exploratory analyses. No correction was applied for the reproduction analyses, as none was used in the original study. For all analyses, a p-value of 0.05 was considered statistically significant. In the figures below, significance levels are denoted as * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

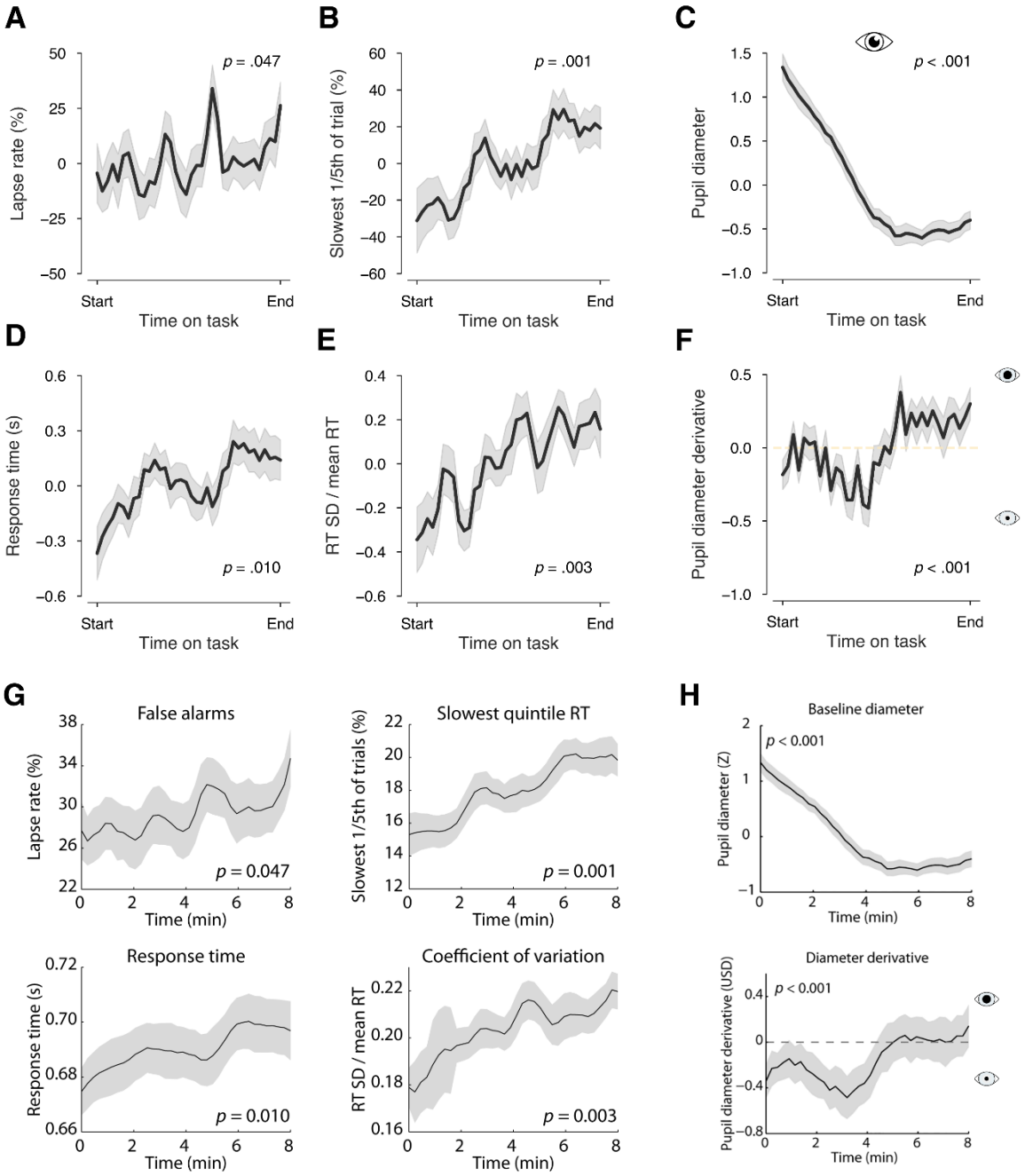
Reproduction Results

Pre-processed time series were Z-scored for each participant and block, after which a straight line was fitted to each series. The slopes of the fitted lines indicated whether the time series was changing over time. These slopes were subsequently averaged across blocks per participant. Using this approach, the time-on-task effects on behavioral and pupillary measures were precisely reproduced, with test statistics identical to those reported in the original study (Figure 4). This also applied to the relationships between behavioral and pupillary measures prior to controlling for time-on-task. The time-controlled outcomes, however, diverged from the original findings. Using the ±15% (test statistics) and ±0.05 (*p*-value) criterion (Miske et al., 2026), quadratic effects of baseline diameter were approximately reproduced for false alarm rate and RTCV, but not for mean RT or the slowest quintile. For linear effects of the diameter derivative, relationships with the slowest quintile and RTCV were approximately reproduced, whereas those with false alarm rate and mean RT were not. A full overview of the differences is provided in Table 2.

The failure to reproduce certain findings likely reflects a process reproducibility failure, which occurs when the results cannot be repeated due to the absence of shared code or the information needed to reconstruct it (Nosek et al., 2022). This means that it was not possible to verify all of the findings by van den Brink et al. (2016). Nevertheless, the successful reproduction of performance decline and pupillary changes over time suggests that the Python-based analytical pipeline closely approximates large parts of the original approach. As the design of the replication study was different from the design of the original study, parts of the code (such as block averaging) were adapted to take into account these differences. Given that the replication study differed in design from the original, some aspects of the code were adapted accordingly. Overall, the reproduction did not provide reason to abandon the analysis plan developed for the replication, which was its primary purpose.

Figure 4

Reproduction of the Behavioral and Pupil Results by van den Brink et al. (2016)



Note. "Start" indicates first window (0 minutes), and "End" the final window (36th window; 8 minutes). Behavioral measures appear to be centered within the available dataset. All measures significantly changed over time in the reproduction. **A.** False alarm rate: $t(27) = 1.74$, $p = .047$. **B.** Slowest quintile: $t(27) = 3.31$, $p = .001$. **C.** Mean RT: $t(27) = 2.49$, $p = .010$. **D.** RTCV: $t(27) = 3.06$, $p = .003$. **E.** Baseline pupil diameter decreased over time ($t(27) = -8.10$, $p < .001$). **F.** The pupil diameter derivative showed a shift from initial constriction to increasing dilation over time ($t(27) = 4.34$, $p < .001$). **G.** Behavioral results reprinted from the original paper © 2016 van den Brink et al. **H.** Pupillary results reprinted from the original paper © 2016 van den Brink et al.

Table 2

Overview of Original (van den Brink et al., 2016) and Reproduced Relationships between Pupillary and Behavioral Measures after Controlling for Time-on-task

Measure	*t*		*p*		Reproduced
	Original	Reproduction	Original	Reproduction	
Quadratic (baseline pupil diameter)					
Lapse rate	1.99	2.11	**.028**	**.022**	Approximately
Mean RT	2.06	1.52	**.025**	.070	No
Slowest quintile	1.45	3.36	.080	**.001**	No
RTCV	2.79	2.62	**.005**	**.007**	Approximately
Linear (pupil diameter derivative)					
Lapse rate	3.09	4.11	**.005**	**< .001**	No
Mean RT	0.93	1.33	.360	.196	No
Slowest quintile	2.13	2.47	**.041**	**.020**	Approximately
RTCV	3.07	2.65	**.005**	**.013**	Approximately

Note. RT = response time; RTCV = response time coefficient of variation. Linear relationships with baseline pupil diameter are not reported, as they were not significant in either of the analyses. Bold p-values indicate statistical significance at $p < .05$.

Replication Results

Performance Differs across Instruction Conditions

Prior to hypothesis testing, psychometric curves (Figure 5.E) were fitted to unsegmented choice data to explore performance across the two instruction conditions (Figure 5.A-C). Lapse rates differed significantly between conditions, whereas the slope ($t(70) = 1.11$, $p = .273$) and the response bias did not ($t(70) = 0.96$, $p = .343$). These results suggest that the differences in lapse rates cannot be attributed to differences in perceptual processing or systematic choice preferences.

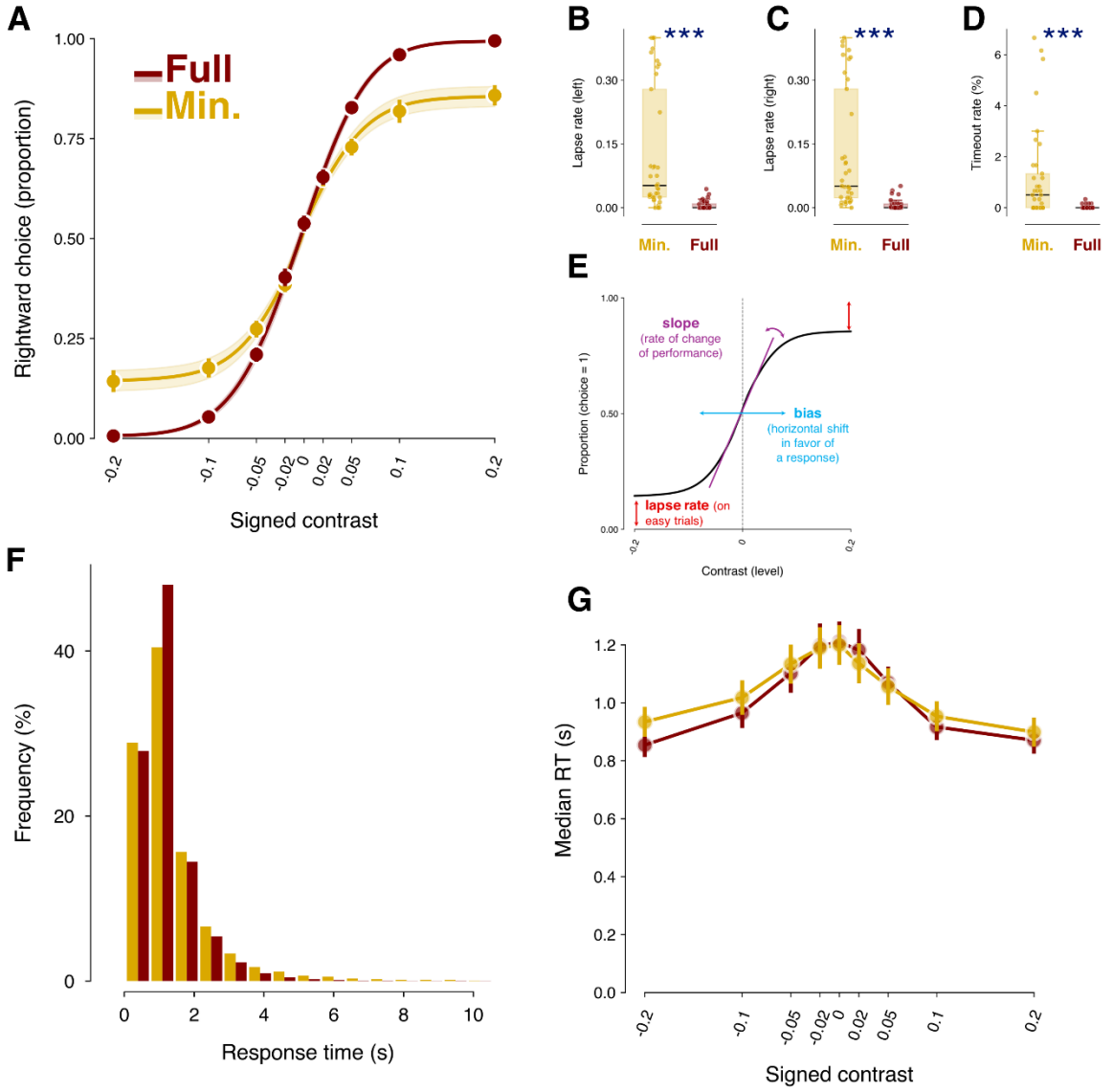
Response time distribution showed a rightward skew (Figure 5.F), consistent with the pattern typically observed in cognitive tasks. Overall mean RT was 1.289s ($SD = 0.991$). Participants in the minimal instructions condition responded at or below this mean on 63.9% of trials, compared to 70.3% in the full instructions condition. Mean RT was slightly longer in the minimal instructions condition ($M = 1.376$s, $SD = 1.155$) than in the full instructions condition ($M = 1.198$s, $SD = 0.773$). To examine how RTs varied across contrast levels, chronometric curves were fitted to the median RT data for each condition separately (Figure 5.G). Two repeated-measures ANOVAs confirmed that RTs differed significantly across contrast values in both groups (minimal instructions: $F(8,288) = 16.99$, $p < .001$; full instructions: $F(8,288) = 47.27$, $p < .001$).

Trials without a recorded response within the time limit were excluded from RT analyses. Fewer than 1% of trials timed out (269 trials, 0.06%), with most occurring in the minimal instructions condition (263 trials, 1.18% of that group's trials) compared to the full instructions condition (6 trials, 0.03%). Consistent with van den Brink et al. (2016), who reported a comparable miss rate of 0.2%,

timed-out trials were not included in hypothesis testing. The difference in time-out rates between the two conditions can be found in Figure 5.D.

Figure 5

Exploratory Performance Metrics across Experimental Conditions



Note. **A.** Average psychometric curve across participants for each experimental condition. Circles indicate mean performance; error bars represent the standard error of the mean (SEM = SD of individual psychometric curves / $\sqrt{N}$). The x-axis shows signed contrast values, which reflect both the location and contrast difference of the target stimulus. Negative values indicate that the target appeared on the left, and positive values indicate that it appeared on the right. The absolute value of each signed contrast reflects the numerical difference in contrast between the target and non-target stimulus. For example, a value of -0.1 indicates that the target was presented on the left at a contrast of 0.60, while the non-target was presented on the right at a contrast of 0.50. **B.** Lapse rates on easy trials with a left-target stimulus, compared between the two conditions ($t(70) = 5.35$, $p < .001$). **C.** Lapse rates on easy trials with a right-target stimulus, compared between the conditions ($t(70) = 5.20$,

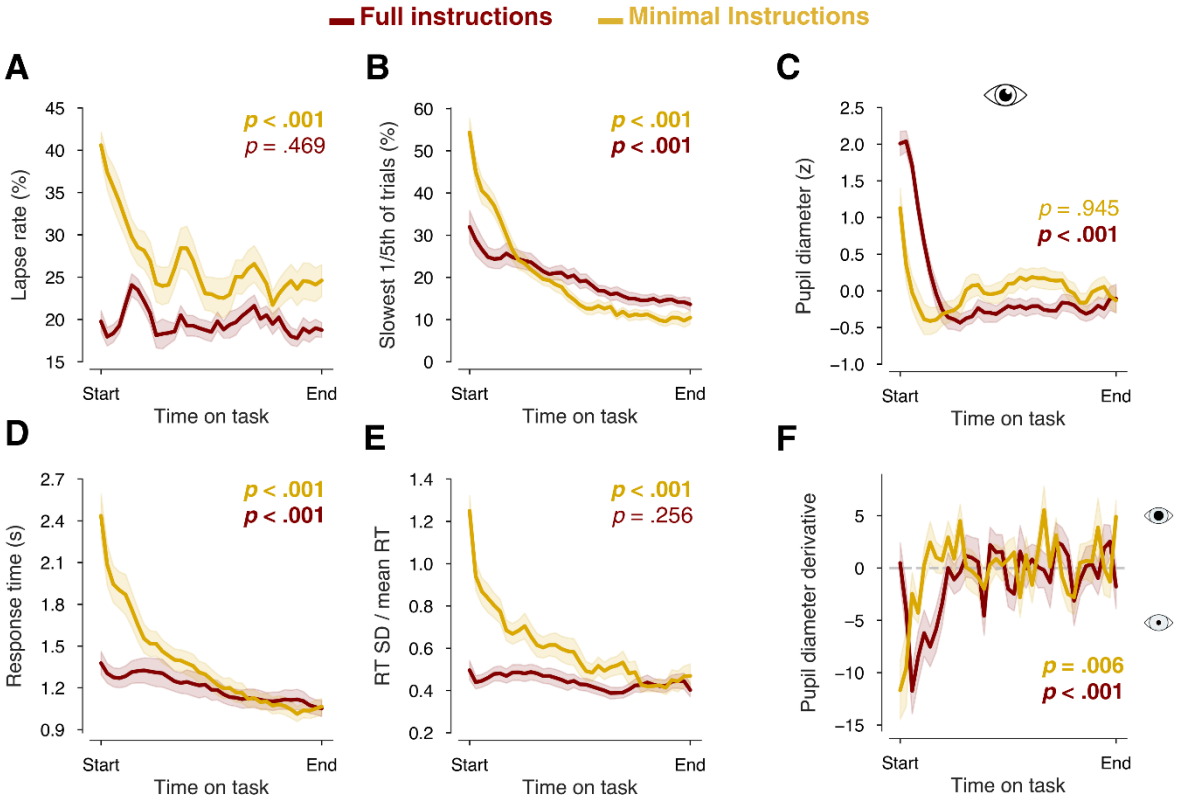
$p < .001$). The psychometric lapse rate captures the proportion of errors on easy trials and is modeled independently from errors on ambiguous trials. **D.** The difference in time-out rates between the two conditions ($t(70) = 3.84$, $p < .001$). **E.** Psychometric curve. Its shape is determined by the slope, reflecting the rate of change in performance; the bias, reflected as a horizontal shift of the curve; and the lapse rates (Wichmann & Hill, 2001). **F.** RT distribution per condition. **G.** Average chronometric curve for each experimental condition. Median RTs did not differ significantly between conditions at any discrete contrast level.

Performance over Time

No Evidence for Performance Decline in the Full Instructions Condition

Figure 6

Behavioral and Pupil Results in the Two Conditions



Note. Annotated p-values are adjusted for multiple comparisons. In panel F, negative values indicate average pupil constriction and positive values indicate dilation. **Full instructions condition.** Confirmatory one-tailed t-tests revealed no significant performance decrements (**A.** false alarm rate: $t(34) = -0.73$, $p_{adj} = .999$; **B.** slowest quintile: $t(34) = -5.09$, $p_{adj} = .999$; **D.** mean RT: $t(34) = -4.75$, $p_{adj} = .999$; **E.** RTCV: $t(34) = -1.37$, $p_{adj} = .999$). Exploratory two-tailed t-tests indicated significant improvements in slowest quintile (**B.** $t(34) = -5.09$, $p_{adj} < .001$, $d_z = 0.86$) and mean RT (**D.** $t(34) = -4.75$, $p_{adj} < .001$, $d_z = 0.80$), but not false alarm rate (**A.** $t(34) = -0.73$, $p_{adj} = .469$, $d_z = 0.12$) or RTCV (**E.** $t(34) = -1.37$, $p_{adj} = .256$, $d_z = 0.23$). Significant time-on-task changes were observed in baseline pupil diameter (**C.** $t(34) = -5.35$, $p_{adj} < .001$, $d_z = 0.90$) and diameter derivative (**F.** $t(34) = 5.89$, $p_{adj} <$

.001, d_z = 0.99). **Minimal instructions condition.** All behavioral measures showed significant improvements over time (**A.** false alarm rate: $t(36) = -7.10$, $p_{adj} < .001$, $d_z = 1.17$; **B.** slowest quintile: $t(36) = -13.57$, $p_{adj} < .001$, $d_z = 2.23$; **D.** mean RT: $t(36) = -13.80$, $p_{adj} < .001$, $d_z = 2.27$; **E.** RTCV: $t(36) = -11.25$, $p_{adj} < .001$, $d_z = 1.85$). There were no observed changes in baseline pupil diameter (**C.** $t(36) = 0.07$, $p_{adj} = .945$, $d_z = 0.01$), whereas the pupil diameter derivative stabilized over time (**F.** $t(36) = 3.31$, $p_{adj} = .006$, $d_z = 0.54$).

To assess whether performance degraded over the course of the task, slope distributions were compared to zero using one-tailed t-tests. Each slope reflected the direction and magnitude of change over time for a given measure, with positive values corresponding to increases in false alarms, proportion of slowest trials, response times, and RTCV, and negative values indicating improvements. Contrary to the hypotheses and the findings by van den Brink et al. (2016), no significant performance decrements were observed in the full instructions condition (Figure 6)

Based on visual inspection of the data, exploratory two-tailed t-tests were conducted to examine whether any directional change occurred irrespective of direction. Several performance metrics showed significant improvements over time, with large effect sizes. Specifically, the proportion of slow trials decreased (Figure 6.B), and the RTs became shorter (Figure 6.D). False alarm rate, the primary measure of performance, did not show significant change, nor did the RTCV. These post-hoc analyses suggest that the observed pattern reflects an effect opposite to that of van den Brink et al. (2016), with performance improving rather than degrading over time.

Performance Improvements in the Minimal Instructions Condition

The changes in performance over time in the minimal instructions condition were explored by comparing the slope distributions to zero using two-tailed t-tests. The two-tailed tests were chosen due to the absence of directional hypotheses during exploratory analyses. Significant improvements were observed across all behavioral measures in the minimal instructions condition (Figure 6.A-B, 6.D-E). False alarm rates decreased over time, responses became faster and less variable (all *p*s < .001). Given the relatively high number of timeouts in this condition, I also examined the changes in the timeout rate over time, which significantly decreased ($t(21) = -7.49$, $p_{adj} < .001$).

Tonic Pupil Fluctuations over Time

Pupillary Measures Show Time-on-task Effects in the Full Instructions Condition

Despite the lack of evidence in support of the first hypothesis, slope distributions of pupillary measures were compared to zero using two-tailed t-tests to examine whether pupillary changes occurred independent of the behavioral findings. In line with the second hypothesis, significant changes were observed in both pupil diameter and its derivative in the full instructions condition. Baseline diameter became smaller over the course of the task (Figure 6.C), while the derivative shifted from increasingly negative values toward positive ones, reflecting a gradual stabilization of diameter (Figure 6.F).

Although the two measures were significantly negatively correlated across windows ($t(34) = -3.86$, $p < .001$), their mean correlation was small ($r = -0.11$, $R^2 = 1.11\%$), indicating their ability to capture largely independent sources of variance in behavior. This finding was comparable to that of van den Brink et al. (2016) and allows for further analyses of the relationship between pupil and behavior.

Changes in the Derivative, but not Baseline Diameter in the Minimal Instructions Condition

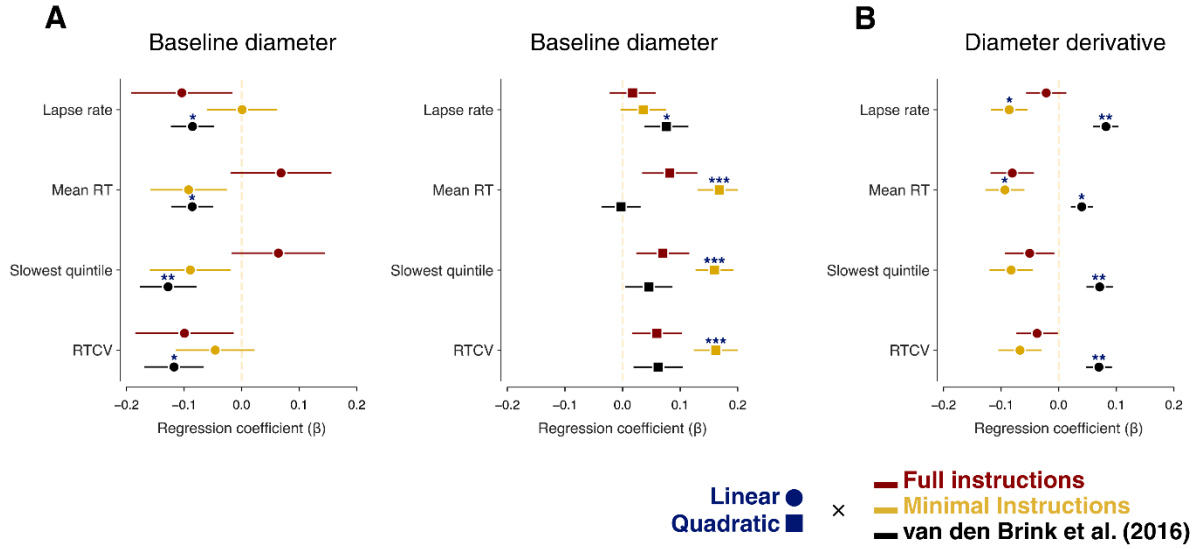
Pupillary fluctuations in the minimal instructions condition were explored by comparing slope distributions of the baseline diameter and its derivative to zero using two-tailed t-tests (Figure 6.C). There were no significant time-related changes observed in the baseline pupil diameter in the minimal instructions condition. However, the diameter derivative was on average negative at the start of the task and increased over time (Figure 6.F). In this condition, the two pupillary measures were also significantly negatively correlated across windows ($t(36) = -2.08$, $p = 0.49$) with a small mean correlation ($r = -0.05$, $R^2 = 0.29\%$).

Tonic Pupil Fluctuations and Task Performance

Multiple regression was used to examine the relationship between Z-scored behavioral and pupillary measures within participants. Following van den Brink et al. (2016), both linear and quadratic regressors were included for baseline diameter, and a linear regressor for the diameter derivative. As before, regression coefficients were compared to zero using t-tests. For the full instructions condition, one-tailed t-tests were used to test the relationship between behavior and baseline diameter, and two-tailed t-tests were used to explore the relationship between behavior and the derivative. In the minimal instructions condition, two-tailed t-tests were used for all comparisons.

Figure 7

The Relationship between Pupil Diameter and Behavior in the Two Conditions and van den Brink et al. (2016) before Controlling for Time-on-task



Note. **A.** No significant linear relationships (left) were observed between baseline pupil diameter and any behavioral measure in either the full instructions condition (lapse rate: $t(34) = -1.18$, $p_{adj} = .253$, d_z

= 0.20; mean RT: $t(34) = 0.78$, $p_{adj} = .781$, $d_z = 0.13$; slowest quintile: $t(34) = 0.79$, $p_{adj} = .781$, $d_z = 0.13$; RTCV: $t(34) = -1.16$, $p_{adj} = .253$, $d_z = 0.20$) or the minimal instructions condition (lapse rate: $t(36) = 0.01$, $p_{adj} = .991$, $d_z = 0.00$; mean RT: $t(36) = -1.39$, $p_{adj} = .260$, $d_z = 0.23$; slowest quintile: $t(36) = -1.28$, $p_{adj} = .281$, $d_z = 0.21$; RTCV: $t(36) = -0.67$, $p_{adj} = .551$, $d_z = 0.11$). Significant linear relationships were observed for all behavioral measures in the van den Brink et al. (2017) dataset. No significant quadratic relationships (right) were observed in the full instructions condition (lapse rate: $t(34) = 0.44$, $p_{adj} = .446$, 0.07; mean RT: $t(34) = 1.71$, $p_{adj} = .253$, $d_z = 0.29$; slowest quintile: $t(34) = 1.51$, $p_{adj} = .253$, $d_z = 0.26$; RTCV: $t(34) = 1.37$, $p_{adj} = .253$, $d_z = 0.23$). Significant quadratic relationships (right) were observed for response time measures in the minimal instructions condition, whereas in the van den Brink et al. (2017) dataset significant quadratic relationships were observed for the false alarm rate measure only. **B.** No significant linear relationships were observed between the diameter derivative and behavioral performance in the full instructions condition (false alarm rate: $t(34) = -0.63$, $p_{adj} = .642$, $d_z = 0.11$; mean RT: $t(34) = -2.14$, $p_{adj} = .253$, $d_z = 0.36$; slowest quintile: $t(34) = -1.19$, $p_{adj} = .419$, $d_z = 0.20$; RTCV: $t(34) = -1.05$, $p_{adj} = .444$, $d_z = 0.18$). Significant relationships were observed in the minimal instructions condition with the false alarm rate ($t(36) = -2.72$, $p_{adj} = .024$, $d_z = 0.45$) and mean RT ($t(36) = -2.75$, $p_{adj} = .024$, $d_z = 0.45$), but not with proportion of slowest RTs ($t(36) = -2.21$, $p_{adj} = .067$, $d_z = 0.36$) and RTCV ($t(36) = -1.78$, $p_{adj} = .143$, $d_z = 0.29$). In the van den Brink et al. (2017) dataset, significant linear relationships were observed between all behavioral measures and the diameter derivative.

Before Controlling for Time-on-Task

No evidence for inverted-U pupil–behavior relationship in the full instructions condition. Contrary to my predictions, no significant linear relationships were observed between behavioral and pupillary measures (Figure 7.A). Baseline pupil diameter was not smaller during windows characterized by more false alarms, slower response times, or greater RT variability, providing no support for a linear arousal–performance relationship. Quadratic relationships between behavioral measures and the baseline diameter were not observed either. This finding does not find support for the proposed Yerkes-Dodson relationship between arousal and performance. Finally, the diameter derivative did not significantly predict behavioral performance (Figure 7.B), failing to replicate the finding of van den Brink et al. (2016) in which periods of pupil stability were associated with poorer performance.

Some evidence for inverted-U pupil–behavior relationship in the minimal instructions condition. In the minimal instructions condition, no significant linear relationships were found between baseline pupil diameter and behavioral performance (Figure 7.A). However, significant positive quadratic relationships were found between baseline diameter and response time measures (Figure 7.A), indicating that RTs became slower and more variable at both extremes of baseline pupil

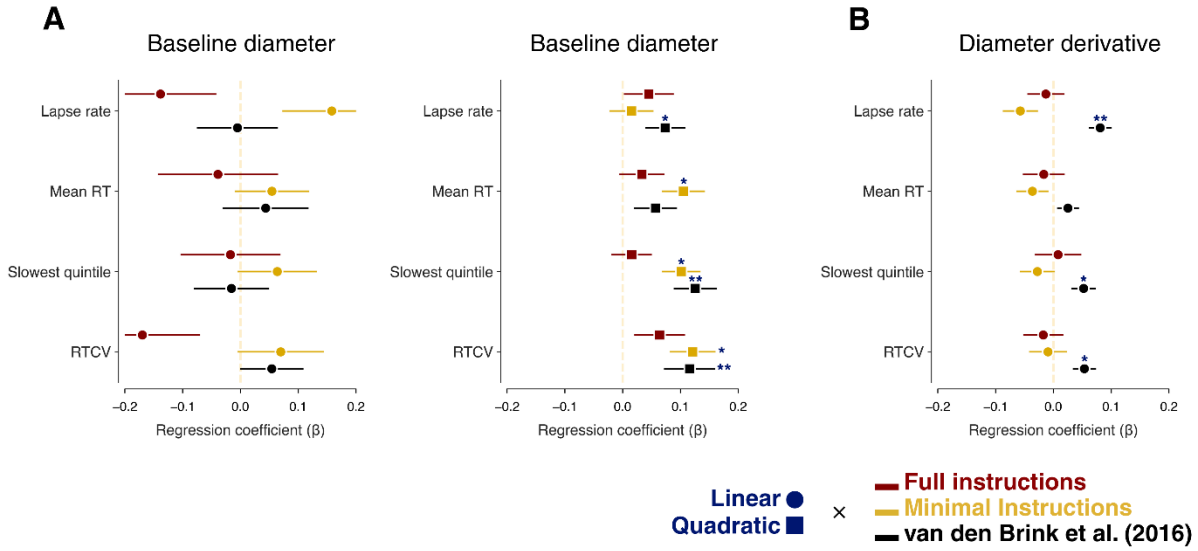
diameter (mean RT: $t(36) = 4.50$, $p_{adj} < .001$, $d_z = 0.74$; slowest quintile: $t(36) = 4.84$, $p_{adj} < .001$, $d_z = 0.80$; RTCV: $t(36) = 4.24$, $p_{adj} < .001$, $d_z = 0.70$). The quadratic relationship was not observed for the false alarm rate ($t(36) = 0.91$, $p_{adj} = .444$, $d_z = 0.15$). Additionally, the diameter derivative showed significant negative linear relationships with false alarm rate and mean RT suggesting that periods of greater pupil stability were associated with faster responses (Figure 7.B).

After Controlling for Time-on-Task

Although performance improved rather than declined over time in the analyses above, additional regressions were conducted to control for its potential confounding effects on the pupil-behavior relationship. To keep the analysis pipeline as close as possible to van den Brink et al. (2016), a time-on-task predictor was included. For each window, this predictor reflected the time point at which a participant completed the last trial in that window. Unlike in the original study, the predictor therefore varied across participants, allowing us to account for individual differences in pacing across windows. The resulting regression coefficients thus reflect relationships that are independent of time-on-task effects.

Figure 8

The Relationship between Pupil Diameter and Behavior in the Two Conditions and van den Brink et al. (2016) after Controlling for Time-on-task

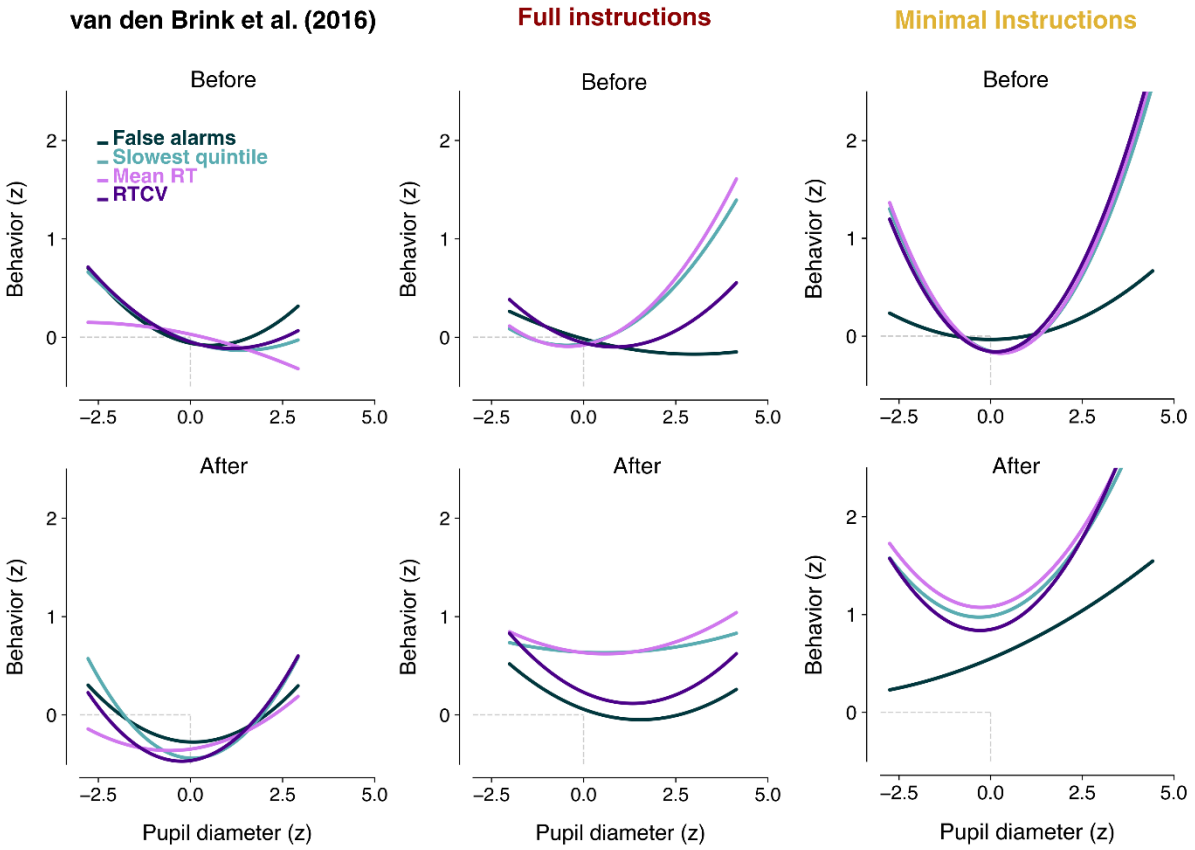


Note. **A.** No significant linear relationships (left) were observed between baseline pupil diameter and any behavioral measure in either the full instructions condition (lapse rate: $t(34) = -1.39$, $p_{adj} = .958$, $d_z = 0.23$; mean RT: $t(34) = -0.45$, $p_{adj} = .925$, $d_z = 0.08$; slowest quintile: $t(34) = -0.30$, $p_{adj} = .915$, $d_z = 0.05$; RTCV: $t(34) = -1.77$, $p_{adj} = .958$, $d_z = 0.30$), the minimal instructions condition (lapse rate: $t(34) = 1.93$, $p_{adj} = .183$, $d_z = 0.32$; mean RT: $t(34) = 0.89$, $p_{adj} = .508$, $d_z = 0.15$; slowest quintile: $t(34) = 0.97$, $p_{adj} = .508$, $d_z = 0.16$; RTCV: $t(34) = 1.03$, $p_{adj} = .508$, $d_z = 0.17$) or the van den Brink et al. (2017) dataset. No significant quadratic relationship (right) were observed in the full instructions

condition (lapse rate: $t(34) = 1.02$, $p_{adj} = .838$, $d_z = 0.17$; mean RT: $t(34) = 0.82$, $p_{adj} = .838$, $d_z = 0.14$; slowest quintile: $t(34) = 0.41$, $p_{adj} = .915$, $d_z = 0.07$; RTCV: $t(34) = 1.45$, $p_{adj} = .838$, $d_z = 0.25$), whereas in the minimal instructions condition all significant quadratic relationships remained (mean RT: $t(36) = 2.75$, $p_{adj} = .037$, $d_z = 0.45$; slowest quintile: $t(36) = 2.97$, $p_{adj} = .032$, $d_z = 0.49$; RTCV: $t(36) = 2.97$, $p_{adj} = .032$, $d_z = 0.49$). In the van den Brink et al. (2017) dataset, quadratic relationships emerged in all but the mean RT measure. This differed from the original publication, which found this relationship significant, but not the slowest quintile (see Table 2). **B.** No significant linear relationships were observed between the diameter derivative and behavioral performance in the full instructions condition (lapse rate: $t(34) = -0.41$, $p_{adj} = .915$, $d_z = 0.07$; mean RT: $t(34) = -0.43$, $p_{adj} = .915$, $d_z = 0.07$; slowest quintile: $t(34) = 0.22$, $p_{adj} = .958$, $d_z = 0.04$; RTCV: $t(34) = -0.49$, $p_{adj} = .915$, $d_z = 0.08$) and the previously observed significant relationships were eliminated in the minimal instructions condition with the inclusion of a time-on-task predictor (lapse rate: $t(36) = -1.74$, $p_{adj} = .217$, $d_z = 0.29$; mean RT: $t(36) = -1.09$, $p_{adj} = .508$, $d_z = 0.18$; slowest quintile: $t(36) = -0.75$, $p_{adj} = .553$, $d_z = 0.12$; RTCV: $t(36) = -0.13$, $p_{adj} = .900$, $d_z = 0.02$). In the van den Brink et al. (2017) dataset, almost all previously observed relationships remained (see Table 2).

Figure 9

Visualization of Quadratic Relationships between Pupil Diameter and Different Measures of Performance Before and After Controlling for Time-on-task



Note. Large values on the y-axis indicate poor behavioral performance (more false alarms, slower and more variable response times), whereas the smaller values indicate better performance.

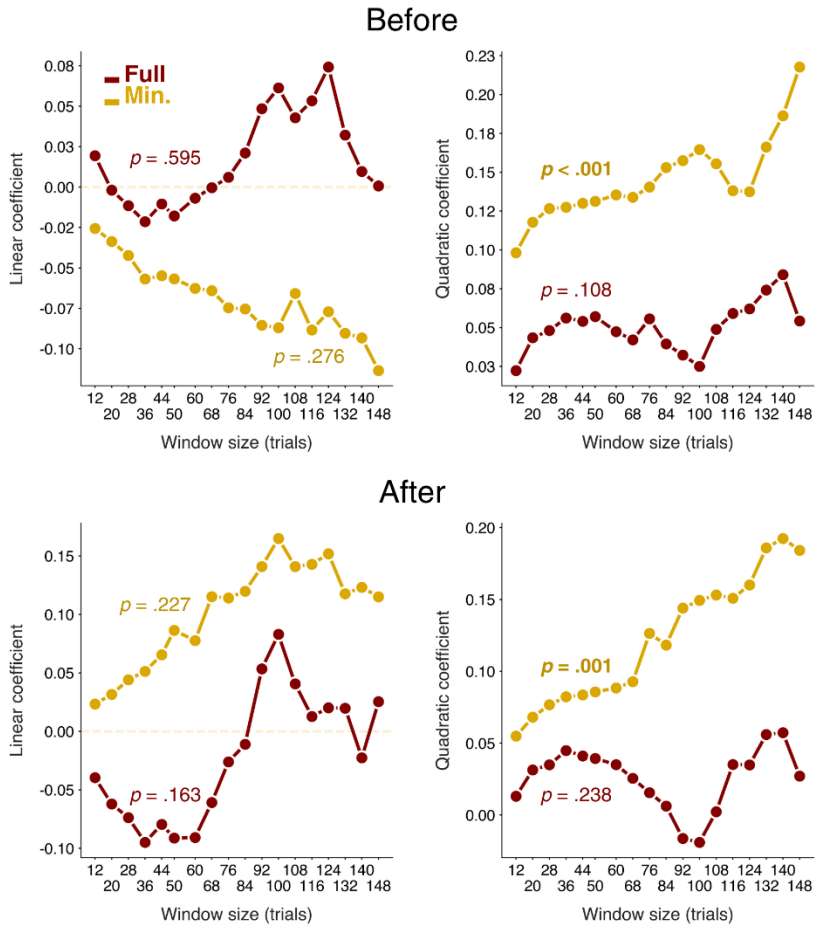
No evidence for time-on-task effects in the full instructions condition. The inclusion of the time-on-task predictor did not alter the previously observed results; all relationships remained non-significant (Figure 8). The time-controlled analyses therefore provided no support for the third hypothesis, which predicted that quadratic relationships would emerge after accounting for time-on-task effects, failing to replicate the original study.

No evidence for time-on-task effects in the minimal instructions condition. In the minimal instructions condition, including the time-on-task predictor rendered the previously observed linear relationships between the pupil derivative and behavioral measures non-significant (all $p > .2$; Figure 8.B). The quadratic relationships between baseline pupil diameter and behavioral measures were, however, preserved (Figure 8.A, Figure 9), suggesting that response time performance is optimal at moderate arousal levels during in the absence of clear task instructions.

Arousal-Performance Relationship Robust in the Minimal, but not in the Full Instructions Condition

Figure 10

Linear and Quadratic AUCs for the Two Conditions Before and After Controlling for Time-on-task



The 50-trial sliding window used in the preprocessing pipeline was adopted from van den Brink et al. (2016), who acknowledged it was an arbitrary choice. To check whether the results depended on this decision, I repeated the regression analyses across window sizes ranging from 12 trials (~40 seconds) to 148 trials (~8 minutes), with step sizes derived proportionally from the original parameters and rounded to the nearest integer. For each window size, linear and quadratic regression coefficients were computed per measure per subject, averaged across behavioral measures, and summarized using the area under the curve (AUC). The AUC measure reflected whether behavioral measures showed a consistent relationship with the two pupil measures overall, independent of the window size choice.

The linear AUC was not significant for either of the two conditions for the baseline pupil diameter before (full instructions: $t(34) = 0.24$, $p = .595$ $d_z = 0.04$; minimal instructions: $t(36) = -1.11$, $p = .276$, $d_z = 1.18$) or after controlling for time-on-task (full instructions: $t(34) = -1.00$, $p = .163$, $d_z = 0.17$; minimal instructions: $t(36) = 1.23$. $p = .227$, $d_z = 0.20$). The linear AUC for the pupil derivative was significant in the minimal instructions condition in both the uncontrolled ($t(36) = -3.19$, $p = .003$, $d_z = 0.52$) and controlled models ($t(36) = -2.48$, $p = .018$, $d_z = 0.41$), whereas in the full instructions condition it was not significant at all (both $p > .4$). In the minimal instructions condition the quadratic AUC for the baseline diameter was significant both before ($t(36) = 4.52$, $p < .001$, $d_z = 0.74$) and after ($t(36) = 3.48$, $p = .001$, $d_z = 0.57$) controlling for time-on-task (Figure 10). This indicated that performance was worst during periods of both relatively low and relatively high arousal, and that this pattern was stable across time scales. In the full instructions condition it was not significant in either model (uncontrolled: $t(34) = 1.26$, $p = .108$, $d_z = 0.21$; controlled: $t(34) = 0.72$, $p = .238$, $d_z = 0.12$), and the direction of the quadratic coefficient reversed across window sizes, suggesting the effect was inconsistent and scale-dependent.

Taken together, this study did not replicate the effects reported by van den Brink et al. (2016) in the main condition, finding no clear link between arousal and performance when full task instructions were given. However, in the minimal instructions condition an inverted-U relationship between arousal and response time measures emerged consistently across all window sizes, independent of the effect of time. This finding suggests that the effect of arousal on performance may depend on the structure and difficulty of the experimental task.

Discussion

This study examined the generalizability of findings by van den Brink et al. (2016) by investigating the effects of time on task engagement using behavioral and physiological markers in a perceptual discrimination task. Contrary to expectations, performance improved rather than degraded over time. Pupil diameter decreased progressively in the full instructions condition but not the minimal instructions condition, while the diameter derivative shifted from initial constriction toward increasing dilation in both conditions - consistent with previous findings. Notably, the hypothesized

linear relationships between behavioral and pupillary measures were not replicated. Instead, response times were longer and more variable when pupil diameter was either relatively small or relatively large, revealing a U-shaped relationship in the minimal instructions condition, but not the full instructions condition. This pattern persisted after controlling for time-on-task and across time scales, suggesting that time did not obscure the true relationship between performance and pupil diameter as observed in van den Brink et al. (2016). In the full instructions condition, however, the relationship between arousal and performance was inconsistent across window sizes, suggesting that the current model failed to capture all factors that contributed to variability in engagement.

Effects of Time on Performance

Prior research on time-on-task effects has consistently shown that performance declines over time in a wide range of paradigms. In studies utilizing discrimination and/or detection tasks without additional manipulations (e.g., rewards, stimuli strength, etc.), performance has been shown to deteriorate over time (Esterman et al., 2013; Hopstaken et al., Jun et al., 2019; 2015b; Mangin et al., 2022; Möckel et al., 2015; Unsworth et al., 2010; Unsworth & Robison, 2016). This pattern, however, alters following the introduction of reward manipulations mid-task, which has been shown to rapidly restore performance (Hopstaken et al., 2015a, 2015b). Task difficulty further moderates temporal patterns, with performance declining on easier tasks but improving on harder ones, as reflected in shorter response times (Hopstaken et al., 2015a). Thus, the manipulation of difficulty via stimulus strength (contrast levels) included in the present study may have mitigated the potential time-related effects. While confirmation of this hypothesis would require further analyses of models that control for differences in stimulus strength, it could indicate that temporal effects on performance (i.e., fatigue effects) could be alleviated by experimental designs that interleave trials of varying difficulty.

Performance on decision-making tasks does not only depend on the observable environment, such as visual differences in stimuli. It also depends on the history of previous choices and stimuli, as even trained observers show substantial history dependence (Frund et al., 2014). Unlike most prior studies (e.g., Beerendonk et al., 2024; Hopstaken et al., 2015b; van den Brink et al., 2016), the present study introduced an additional level of manipulation: block-wise biases in target location probability (see Figure 2.C). The analytical pipeline of van den Brink et al. (2016) did not allow for control of such biases (or priors), meaning the sliding window approach used here aggregated trials with varying levels of bias. This likely introduced unexpected noise into the analyses. While sliding window and binning approaches are common (Beerendonk et al., 2024; van den Brink et al., 2016), they may not be optimal for designs with this kind of manipulation. Future research should therefore explore analytical approaches that better account for historical dependencies and afford greater statistical control. Several such approaches have been developed, notably the input-output hidden Markov models (Ashwood et al., 2022) and models based on artificial neural networks (Eckstein et al., 2023; Urai, 2026).

A further methodological difference that may account for the present findings is the use of a non-continuous paradigm. Unlike van den Brink et al. (2016), the present study used a task with clearly separated trials, each beginning and ending with a fixation dot (see Figure 2.B). Though both continuous (Esterman et al., 2013; Jun et al., 2019) and non-continuous (Hopstaken et al., 2015a) paradigms appear in the literature, these designs may not capture identical processes. Specifically, inter-trial intervals in non-continuous tasks may function as micro-breaks, allowing for partial fatigue recovery. Consistent with this, some work has shown that micro-breaks can stabilize performance (Dianita et al., 2024; Sharpe et al., 2025), which could account for the decrease in response time variability observed in the present study. However, this line of research remains relatively novel and is largely confined to applied settings, meaning its relevance to perceptual decision-making tasks warrants further investigation.

Effects of Time on the Pupil

Despite the divergence in behavioral findings, pupillary dynamics were partially replicated. In the full instructions condition, both measures of baseline diameter mirrored the original pattern (van den Brink et al., 2016), whereas in the minimal instructions condition, only the time-dependent stabilization of the diameter (diameter derivative) was observed. Declines in baseline pupil diameter over time have been reported across a range of studies, regardless of whether performance decrements were present (Hopstaken et al., 2015a; Pielage et al., 2021; van den Brink et al., 2016), though not universally (Hopstaken et al., 2015b). According to the LC-NE theory of arousal (Aston-Jones & Cohen, 2005), such declines reflect a fatigue-driven shift away from intermediate, optimal pupil sizes following prolonged attentional effort. The findings of the present study suggest that these physiological dynamics may act independently of task performance in the full instructions condition. In the minimal instructions condition, the pupil diameter initially declined before progressively increasing over the course of the task. This pattern may reflect the greater cognitive demands imposed by ambiguous instructions, sustaining or recovering arousal over time. Consistent with this, some work suggests that challenging tasks can drive performance-related increases in arousal (Sayalı et al., 2023), which may explain the concurrent improvement in performance observed in this condition.

Relationship between Performance and Arousal

The present study found no evidence that larger baseline pupil sizes relate to performance improvements (cf. Podvalny et al., 2021; van den Brink et al., 2016). Instead, performance appeared optimal at intermediate pupil diameters, consistent with a quadratic relationship. This pattern has been previously observed, with several studies finding a quadratic relationship between performance and pupil diameter both before and after controlling for the effect of time (Beerendonk et al. 2024; Van Kempen et al., 2019). Other studies have found mixed results, with some performance measures showing linear and others quadratic relationships with pupil diameter (Podvalny et al., 2021).

These findings align with the broader principle that arousal facilitates performance on easy tasks but follows an inverted-U relationship on more difficult ones (Sörensen et al., 2022; Yerkes &

Dodson, 1908). This may also explain the absence of quadratic relationships in van den Brink et al. (2016) prior to controlling for time - if their task was sufficiently easy, only the ascending portion of the U-shaped curve may have been captured, producing an apparently linear relationship. The greater difficulty of the present task may have engaged a wider range of the arousal spectrum, allowing the full quadratic relationship to emerge. This could also extend to the condition differences observed here: the minimal instructions condition, being more cognitively demanding, may have better sampled the arousal range, producing stronger quadratic relationships that were less susceptible to the effects of time.

Finally, since the task used in this study was originally developed to enable cross-species comparison with recent findings on task engagement in mice (Ashwood et al., 2022; Hulsey et al., 2024; Johnson et al., 2025), the present results raise a question about how the minimal instructions condition in particular maps onto animal data. As mice acquire task structure through experience rather than explicit instruction, the minimal instructions condition may represent a closer human analogue to rodent learning paradigms. Therefore, future research could use this paradigm to investigate whether the performance-arousal relationship in humans under minimal instruction parallel those observed in rodents, offering a bridge between human and rodent decision-making.

Limitations and Future Directions

The present study adopted a threefold approach to explore the temporal dynamics of, and relationship between, performance and arousal during sustained attention. First, a computational reproduction of van den Brink et al. (2016) was conducted, followed by a conceptual replication aimed at establishing the generalizability of their findings. Finally, the replication was extended to a manipulated condition to illustrate differences between the two. The computational reproduction revealed that some, but not all, findings emerge when the same dataset is analysed with different code. Albeit small (11% of statistics did not reproduce) and based on a single study, this discrepancy highlights the importance of transparent and open code sharing for the success of replication research in psychology and science more broadly (Breznau et al., 2025; Laurinavichyute et al., 2022). It also limits the conclusions that can be drawn from the present replication.

Beyond reproducibility concerns, several analytical decisions further constrain interpretation. The data inclusion criteria were finalised following post-hoc quality control, and, although the exclusion thresholds were conservative and largely consistent with the original proposal, future studies would benefit from clearly specified pre-registration plans to minimise potential bias (Bakker et al., 2020). Additionally, the sliding window procedure employed here may not be well-suited to the present paradigm, given the multiple simultaneous manipulations within the paradigm (e.g., stimulus location, contrast level, frequency of location change, etc.). Considering the somewhat arbitrary nature of the procedural design, future work should consider comparing multiple preprocessing pipelines to determine whether observed effects reflect genuine phenomena or are partly a product of analytic choices.

A further limitation concerns the operationalisation of arousal. The present study relied on the pupil as a proxy of arousal - however, pupil size is not exclusively driven by LC-NE activity but is additionally influenced by other neuromodulatory systems, including dopaminergic, and serotonergic pathways (Cazettes et al., 2021; Grove et al., 2022; Grujic et al., 2024; Joshi & Gold, 2020). While work is being carried out to characterise the precise relationship between pupillary and stimulated LC-NE activity (see Weiss et al., 2026), the strength of inference in the present work could be improved by incorporating additional arousal measures, such as electroencephalographic markers (e.g., P3 amplitude or alpha power; Kopčanová et al., 2025; Murphy et al., 2011) or physiological signals such as heart rate or breathing indices (Iwamoto et al., 2023), to provide converging evidence.

Perhaps the most fundamental limitation, however, is that this remains a single study. The literature on arousal and performance is extensive, yet evidence for any general underlying mechanism is sparse. This may be a consequence of the field's emphasis on novel methodologies at the expense of systematic replication. As a result, it remains largely unknown whether findings from any given paradigm generalise beyond that specific design. While studies like Beerendonk et al. (2024), which utilize data from multiple tasks, help illuminate these relationships, they remain bound by their own methodological choices. A valuable next step would therefore be a collaborative effort to re-analyse and compare existing datasets measuring the same constructs using multiple analytical approaches, both in human and non-human populations (e.g., based on efforts by The International Brain Laboratory). If anything, the present study suggests that such efforts may be both worthwhile and feasible - regardless of the employed study designs.

Conclusion

In conclusion, the present study examined the generalizability of van den Brink et al. (2016) by exploring behavioral and physiological dynamics in sustained perceptual decisions. Just as the ripeness of some apples is harder to judge than that of others, the present task varied stimulus contrast to create trials of differing difficulty, unlike the original study where all stimuli were equally difficult. While performance did not degrade but rather improved over time, the pupillary dynamics were partially replicated, and a quadratic rather than linear relationship between arousal and performance emerged independent of time-on-task. These findings shed new light on the boundary conditions of established arousal-performance relationships, and highlight the value of replication studies in building a more complete understanding of the mechanisms underlying perceptual decision-making.

References

Alnaes, D., Sneve, M. H., Espeseth, T., Endestad, T., Van De Pavert, S. H. P., & Laeng, B. (2014). Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus. *Journal of Vision*, *14*(4), 1–1. https://doi.org/10.1167/14.4.1

Ashwood, Z. C., Roy, N. A., Stone, I. R., The International Brain Laboratory, Urai, A. E., Churchland, A. K., Pouget, A., & Pillow, J. W. (2022). Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience*, *25*(2), 201–212. https://doi.org/10.1038/s41593-021-01007-z

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*(1), 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Aston-Jones, G., Rajkowski, J., & Cohen, J. (1999). Role of locus coeruleus in attention and behavioral flexibility. *Biological Psychiatry*, *46*(9), 1309–1320. https://doi.org/10.1016/S0006-3223(99)00140-7

Aston-Jones, G., & Waterhouse, B. (2016). Locus coeruleus: From global projection system to adaptive regulation of behavior. *Brain Research*, *1645*, 75–78. https://doi.org/10.1016/j.brainres.2016.03.001

Avitan, L., & Stringer, C. (2022). Not so spontaneous: Multi-dimensional representations of behaviors and context in sensory areas. *Neuron*, *110*(19), 3064–3075. https://doi.org/10.1016/j.neuron.2022.06.019

Awwad Shiekh Hasan, B., Joosten, E., & Neri, P. (2012). Estimation of internal noise using double passes: Does it matter how the second pass is delivered? *Vision Research*, *69*, 1–9. https://doi.org/10.1016/j.visres.2012.06.014

Ayasse, N. D., & Wingfield, A. (2020). Anticipatory Baseline Pupil Diameter Is Sensitive to Differences in Hearing Thresholds. *Frontiers in Psychology*, *10*, 2947. https://doi.org/10.3389/fpsyg.2019.02947

Bakker, M., Veldkamp, C. L. S., Van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., & Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, *18*(12), e3000937. https://doi.org/10.1371/journal.pbio.3000937

Beatty, J. (1982a). Phasic Not Tonic Pupillary Responses Vary With Auditory Vigilance Performance. *Psychophysiology*, *19*(2), 167–172. https://doi.org/10.1111/j.1469-8986.1982.tb02540.x

Beatty, J. (1982b). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276–292. https://doi.org/10.1037/0033-2909.91.2.276

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability. *Neuron*, *74*(1), 30–39. https://doi.org/10.1016/j.neuron.2012.03.016

Beerendonk, L., Mejías, J. F., Nuiten, S. A., De Gee, J. W., Fahrenfort, J. J., & Van Gaal, S. (2024). A disinhibitory circuit mechanism explains a general principle of peak performance during mid-level arousal. *Proceedings of the National Academy of Sciences*, *121*(5), e2312898121. https://doi.org/10.1073/pnas.2312898121

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., Krypotos, A.-M., Lerche, V., Logan, G. D., Palmeri, T. J., Van Ravenzwaaij, D., Servant, M., Singmann, H., Starns, J. J., Voss, A., Wiecki, T. V., Matzke, D., & Wagenmakers, E.-J. (2018). Estimating across-trial variability parameters of the Diffusion Decision Model: Expert advice and recommendations. *Journal of Mathematical Psychology*, *87*, 46–75. https://doi.org/10.1016/j.jmp.2018.09.004

Bolkan, S. S., Stone, I. R., Pinto, L., Ashwood, Z. C., Iravedra Garcia, J. M., Herman, A. L., Singh, P., Bandi, A., Cox, J., Zimmerman, C. A., Cho, J. R., Engelhard, B., Pillow, J. W., & Witten, I. B. (2022). Opponent control of behavior by dorsomedial striatal pathways depends on task demands and internal state. *Nature Neuroscience*, *25*(3), 345–357. https://doi.org/10.1038/s41593-022-01021-9

Bouter, L. M., & Riet, G. T. (2021). Replication Research Series-Paper 2: Empirical research must be replicated before its findings can be trusted. *Journal of Clinical Epidemiology*, *129*, 188–190. https://doi.org/10.1016/j.jclinepi.2020.09.032

Breton-Provencher, V., Drummond, G. T., & Sur, M. (2021). Locus Coeruleus Norepinephrine in Learned Behavior: Anatomical Modularity and Spatiotemporal Integration in Targets. *Frontiers in Neural Circuits*, *15*, 638007. https://doi.org/10.3389/fncir.2021.638007

Breznau, N., Rinke, E. M., Wuttke, A., Adem, M., Adriaans, J., Akdeniz, E., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Bai, L., Balzer, D., Bauer, P. C., Bauer, G., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., … Hochman, O. (2025). *The reliability of replications: A study in computational reproductions*. *12*(3). https://doi.org/10.1098/rsos.241038

Bruya, B., & Tang, Y.-Y. (2018). Is Attention Really Effort? Revisiting Daniel Kahneman's Influential 1973 Book Attention and Effort. *Frontiers in Psychology*, *9*, 1133. https://doi.org/10.3389/fpsyg.2018.01133

Cazettes, F., Reato, D., Morais, J. P., Renart, A., & Mainen, Z. F. (2021). Phasic Activation of Dorsal Raphe Serotonergic Neurons Increases Pupil Size. *Current Biology*, *31*(1), 192-197.e4. https://doi.org/10.1016/j.cub.2020.09.090

Charnay, Y., & Leger, L. (2010). Brain serotonergic circuitries. *Dialogues in Clinical Neuroscience*, *12*(4), 471–487. https://doi.org/10.31887/DCNS.2010.12.4/ycharnay

Dianita, O., Kitayama, K., Ueda, K., Ishii, H., Shimoda, H., & Obayashi, F. (2024). Systematic micro-breaks affect concentration during cognitive comparison tasks: Quantitative and qualitative measurements. *Advances in Computational Intelligence*, *4*(3), 7. https://doi.org/10.1007/s43674-024-00074-6

Drugowitsch, J., Wyart, V., Devauchelle, A.-D., & Koechlin, E. (2016). Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron*, *92*(6), 1398–1411. https://doi.org/10.1016/j.neuron.2016.11.005

Duffy, J. S., Bellgrove, M. A., Murphy, P. R., & O'Connell, R. G. (2025). Disentangling sources of variability in decision-making. *Nature Reviews Neuroscience*, *26*(5), 247–262. https://doi.org/10.1038/s41583-025-00916-3

Eckstein, M. K., Summerfield, C., Daw, N. D., & Miller, K. J. (2023). Predictive and Interpretable: Combining Artificial Neural Networks and Classic Cognitive Models to Understand Human Learning and Decision Making. *bioRxiv*. https://doi.org/10.1101/2023.05.17.541226

Enwereuzor, C.U., Johnson, P. A., & Urai, A.E. (2024). *human_IBL_cursor*. github.com/cocosys-lab/human_IBL_cursor

Esterman, M., Noonan, S. K., Rosenberg, M., & DeGutis, J. (2013). In the Zone or Zoning Out? Tracking Behavioral and Neural Fluctuations During Sustained Attention. *Cerebral Cortex*, *23*(11), 2712–2723. https://doi.org/10.1093/cercor/bhs261

Flavell, S. W., Gogolla, N., Lovett-Barron, M., & Zelikowsky, M. (2022). The emergence and influence of internal states. *Neuron*, *110*(16), 2545–2570. https://doi.org/10.1016/j.neuron.2022.04.030

Flavell, S. W., Pokala, N., Macosko, E. Z., Albrecht, D. R., Larsch, J., & Bargmann, C. I. (2013). Serotonin and the Neuropeptide PDF Initiate and Extend Opposing Behavioral States in C. elegans. *Cell*, *154*(5), 1023–1035. https://doi.org/10.1016/j.cell.2013.08.001

Fortenbaugh, F. C., DeGutis, J., & Esterman, M. (2017). Recent theoretical, neural, and clinical advances in sustained attention research. *Annals of the New York Academy of Sciences*, *1396*(1), 70–91. https://doi.org/10.1111/nyas.13318

Frund, I., Wichmann, F. A., & Macke, J. H. (2014). Quantifying the effect of intertrial dependence on perceptual decisions. *Journal of Vision*, *14*(7), 9–9. https://doi.org/10.1167/14.7.9

Gilden, D. L., Thornton, T., & Mallon, M. W. (1995). 1/*f* Noise in Human Cognition. *Science*, *267*(5205), 1837–1839. https://doi.org/10.1126/science.7892611

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

Gold, J. I., & Shadlen, M. N. (2007). The Neural Basis of Decision Making. *Annual Review of Neuroscience*, *30*(1), 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Grandchamp, R., Braboszcz, C., & Delorme, A. (2014). Oculometric variations during mind wandering. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00031

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. In *Signal detection theory and psychophysics. By Green* (p. 40997). https://ui.adsabs.harvard.edu/abs/1966ntrs.book40997G

Greene, A. S., Horien, C., Barson, D., Scheinost, D., & Constable, R. T. (2023). Why is everyone talking about brain state? *Trends in Neurosciences*, *46*(7), 508–524. https://doi.org/10.1016/j.tins.2023.04.001

Grove, J. C. R., Gray, L. A., La Santa Medina, N., Sivakumar, N., Ahn, J. S., Corpuz, T. V., Berke, J. D., Kreitzer, A. C., & Knight, Z. A. (2022). Dopamine subsystems that track internal states. *Nature*, *608*(7922), 374–380. https://doi.org/10.1038/s41586-022-04954-0

Grujic, N., Polania, R., & Burdakov, D. (2024). Neurobehavioral meaning of pupil size. *Neuron*, *112*(20), 3381–3395. https://doi.org/10.1016/j.neuron.2024.05.029

Hebb, D. O. (1955). Drives and the C. N. S. (conceptual nervous system). *Psychological Review*, *62*(4), 243–254. https://doi.org/10.1037/h0041823

Hong, L., Walz, J. M., & Sajda, P. (2014). Your Eyes Give You Away: Prestimulus Changes in Pupil Diameter Correlate with Poststimulus Task-Related EEG Dynamics. *PLoS ONE*, *9*(3), e91321. https://doi.org/10.1371/journal.pone.0091321

Hopstaken, J. F., Van Der Linden, D., Bakker, A. B., & Kompier, M. A. J. (2015a). A multifaceted investigation of the link between mental fatigue and task disengagement. *Psychophysiology*, *52*(3), 305–315. https://doi.org/10.1111/psyp.12339

Hopstaken, J. F., Van Der Linden, D., Bakker, A. B., & Kompier, M. A. J. (2015b). The window of my eyes: Task disengagement and mental fatigue covary with pupil dynamics. *Biological Psychology*, *110*, 100–106. https://doi.org/10.1016/j.biopsycho.2015.06.013

Hulsey, D., Zumwalt, K., Mazzucato, L., McCormick, D. A., & Jaramillo, S. (2024). Decision-making dynamics are predicted by arousal and uninstructed movements. *Cell Reports*, *43*(2), 113709. https://doi.org/10.1016/j.celrep.2024.113709

Iwamoto, M., Yonekura, S., Atsumi, N., Hirabayashi, S., Kanazawa, H., & Kuniyoshi, Y. (2023). Respiratory entrainment of the locus coeruleus modulates arousal level to avoid physical risks from external vibration. *Scientific Reports*, *13*(1), 7069. https://doi.org/10.1038/s41598-023-32995-6

Jepma, M., & Nieuwenhuis, S. (2011). Pupil Diameter Predicts Changes in the Exploration–Exploitation Trade-off: Evidence for the Adaptive Gain Theory. *Journal of Cognitive Neuroscience*, *23*(7), 1587–1596. https://doi.org/10.1162/jocn.2010.21548

Johnson, P. A., Nieuwenhuis, S., Mejías, J., & Urai, A. E. (2025). A dynamical systems model of arousal-driven behavioural state transitions. *bioRxiv*, 2025.10.31.685593. https://doi.org/10.1101/2025.10.31.685593

Johnson, P. A., Urai, A. E., Frach, A., & Enwereuzor, C. U. (2026). *Human_ibl_snapshots*. github.com/anne-urai/human_ibl_snapshots

Joshi, S., & Gold, J. I. (2020). Pupil Size as a Window on Neural Substrates of Cognition. *Trends in Cognitive Sciences*, *24*(6), 466–480. https://doi.org/10.1016/j.tics.2020.03.005

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

Jun, J., Remington, R. W., Koutstaal, W., & Jiang, Y. V. (2019). Characteristics of sustaining attention in a gradual-onset continuous performance task. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(3), 386–401. https://doi.org/10.1037/xhp0000604

Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science*, *154*(3756), 1583–1585. https://doi.org/10.1126/science.154.3756.1583

Klein, M. O., Battagello, D. S., Cardoso, A. R., Hauser, D. N., Bittencourt, J. C., & Correa, R. G. (2019). Dopamine: Functions, Signaling, and Association with Neurological Diseases. *Cellular and Molecular Neurobiology*, *39*(1), 31–59. https://doi.org/10.1007/s10571-018-0632-3

Kopčanová, M., Thut, G., Benwell, C. S. Y., & Keitel, C. (2025). Characterising time-on-task effects on oscillatory and aperiodic EEG components and their co-variation with visual task performance. *Imaging Neuroscience*, *3*, imag_a_00566. https://doi.org/10.1162/imag_a_00566

Korbmacher, M., Azevedo, F., Pennington, C. R., Hartmann, H., Pownall, M., Schmidt, K., Elsherif, M., Breznau, N., Robertson, O., Kalandadze, T., Yu, S., Baker, B. J., O'Mahony, A., Olsnes, J. Ø.-S., Shaw, J. J., Gjoneska, B., Yamada, Y., Röer, J. P., Murphy, J., … Evans, T. (2023). The replication crisis has led to positive structural, procedural, and community changes. *Communications Psychology*, *1*(1), 3. https://doi.org/10.1038/s44271-023-00003-2

Krafczyk, M. S., Shi, A., Bhaskar, A., Marinov, D., & Stodden, V. (2021). Learning from reproducing computational results: Introducing three principles and the *Reproduction Package*. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2197), rsta.2020.0069, 20200069. https://doi.org/10.1098/rsta.2020.0069

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Langner, R., & Eickhoff, S. B. (2013). Sustaining attention to simple tasks: A meta-analytic review of the neural mechanisms of vigilant attention. *Psychological Bulletin*, *139*(4), 870–900. https://doi.org/10.1037/a0030694

Larsen, R. S., & Waters, J. (2018). Neuromodulatory Correlates of Pupil Dilation. *Frontiers in Neural Circuits*, *12*, 21. https://doi.org/10.3389/fncir.2018.00021

Laurinavichyute, A., Yadav, H., & Vasishth, S. (2022). Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy. *Journal of Memory and Language*, *125*, 104332. https://doi.org/10.1016/j.jml.2022.104332

Mangin, T., Audiffren, M., Lorcery, A., Mirabelli, F., Benraiss, A., & André, N. (2022). A plausible link between the time-on-task effect and the sequential task effect. *Frontiers in Psychology*, *13*, 998393. https://doi.org/10.3389/fpsyg.2022.998393

Martin, J. T., Whittaker, A. H., & Johnston, S. J. (2022). Pupillometry and the vigilance decrement: Task-evoked but not baseline pupil measures reflect declining performance in visual vigilance tasks. *European Journal of Neuroscience*, *55*(3), 778–799. https://doi.org/10.1111/ejn.15585

McCormick, D. A., Nestvogel, D. B., & He, B. J. (2020). Neuromodulation of Brain State and Behavior. *Annual Review of Neuroscience*, *43*(1), 391–415. https://doi.org/10.1146/annurev-neuro-100219-105424

McGinley, M. J., David, S. V., & McCormick, D. A. (2015). Cortical Membrane Potential Signature of Optimal States for Sensory Signal Detection. *Neuron*, *87*(1), 179–192. https://doi.org/10.1016/j.neuron.2015.05.038

McGinley, M. J., Vinck, M., Reimer, J., Batista-Brito, R., Zagha, E., Cadwell, C. R., Tolias, A. S., Cardin, J. A., & McCormick, D. A. (2015). Waking State: Rapid Variations Modulate Neural and Behavioral Responses. *Neuron*, *87*(6), 1143–1161. https://doi.org/10.1016/j.neuron.2015.09.012

McLaughlin, D. J., Zink, M. E., Gaunt, L., Reilly, J., Sommers, M. S., Van Engen, K. J., & Peelle, J. E. (2023). Give me a break! Unavoidable fatigue effects in cognitive pupillometry. *Psychophysiology*, *60*(7), e14256. https://doi.org/10.1111/psyp.14256

Miske, O., Abatayo, A. L., Daley, M., Dirzo, M., Fox, N., Haber, N., Hahn, K. M., Struhl, M. K., Mawhinney, B., Silverstein, P., Stankov, T., Tyner, A. H., Adamkovič, M., Alzahawi, S., Anafinova, S., Awtrey, E., Axxe, E., Bailey, J., Bakker, B. N., … Errington, T. M. (2026). Investigating the reproducibility of the social and behavioural sciences. *Nature*, *652*(8108), 126–134. https://doi.org/10.1038/s41586-026-10203-5

Möckel, T., Beste, C., & Wascher, E. (2015). The Effects of Time on Task in Response Selection—An ERP Study of Mental Fatigue. *Scientific Reports*, *5*(1), 10113. https://doi.org/10.1038/srep10113

Montefusco-Siegmund, R., Schwalm, M., Rosales Jubal, E., Devia, C., Egaña, J. I., & Maldonado, P. E. (2022). Alpha EEG Activity and Pupil Diameter Coupling during Inactive Wakefulness in Humans. *Eneuro*, *9*(2), ENEURO.0060-21.2022. https://doi.org/10.1523/ENEURO.0060-21.2022

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012). Bias in the Brain: A Diffusion Model Analysis of Prior Probability and Potential Payoff. *The Journal of Neuroscience*, *32*(7), 2335–2343. https://doi.org/10.1523/JNEUROSCI.4156-11.2012

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

Newsome, W. T., Britten, K. H., & Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. *Nature*, *341*(6237), 52–54. https://doi.org/10.1038/341052a0

Nienborg, H., & Cumming, B. G. (2009). Decision-related activity in sensory neurons reflects more than a neuron's causal effect. *Nature*, *459*(7243), 89–92. https://doi.org/10.1038/nature07821

Nieuwenhuis, S. (2024). Arousal and performance: Revisiting the famous inverted-U-shaped curve. *Trends in Cognitive Sciences*, *28*(5), 394–396. https://doi.org/10.1016/j.tics.2024.03.011

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus—Norepinephrine system. *Psychological Bulletin*, *131*(4), 510–532. https://doi.org/10.1037/0033-2909.131.4.510

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, *73*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Oken, B. S., Salinsky, M. C., & Elsas, S. M. (2006). Vigilance, alertness, or sustained attention: Physiological basis and measurement. *Clinical Neurophysiology*, *117*(9), 1885–1901. https://doi.org/10.1016/j.clinph.2006.01.017

Parsons, S., Azevedo, F., Elsherif, M. M., Guay, S., Shahim, O. N., Govaart, G. H., Norris, E., O'Mahony, A., Parker, A. J., Todorovic, A., Pennington, C. R., Garcia-Pelegrin, E., Lazić, A., Robertson, O., Middleton, S. L., Valentini, B., McCuaig, J., Baker, B. J., Collins, E., … Aczel, B. (2022). A community-sourced glossary of open scholarship terms. *Nature Human Behaviour*, *6*(3), 312–318. https://doi.org/10.1038/s41562-021-01269-4

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Perquin, M. N., Heed, T., & Kayser, C. (2024). Variance (Un)Explained: Experimental Conditions and Temporal Dependencies Explain Similarly Small Proportions of Reaction Time Variability in Linear Models of Perceptual and Cognitive Tasks. *Journal of Experimental Psychology: General*, *153*(12), 3107–3129. https://doi.org/10.1037/xge0001630

Pfaff, D. W., Martin, E. M., & Faber, D. (2012). Origins of arousal: Roles for medullary reticular neurons. *Trends in Neurosciences*, *35*(8), 468–476. https://doi.org/10.1016/j.tins.2012.04.008

Pielage, H., Zekveld, A. A., Saunders, G. H., Versfeld, N. J., Lunner, T., & Kramer, S. E. (2021). The Presence of Another Individual Influences Listening Effort, But Not Performance. *Ear & Hearing*, *42*(6), 1577–1589. https://doi.org/10.1097/AUD.0000000000001046

Podvalny, E., King, L. E., & He, B. J. (2021). Spectral signature and behavioral consequence of spontaneous shifts of pupil-linked arousal in human. *eLife*, *10*, e68265. https://doi.org/10.7554/eLife.68265

Prusky, G. T., & Douglas, R. M. (2004). Characterization of mouse cortical spatial vision. *Vision Research*, *44*(28), 3411–3418. https://doi.org/10.1016/j.visres.2004.09.001

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260–281. https://doi.org/10.1016/j.tics.2016.01.007

Ratcliff, R., Voskuilen, C., & McKoon, G. (2018). Internal and external sources of variability in perceptual decision-making. *Psychological Review*, *125*(1), 33–46. https://doi.org/10.1037/rev0000080

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*(1), 13289. https://doi.org/10.1038/ncomms13289

Roy, N. A., Bak, J. H., Akrami, A., Brody, C. D., & Pillow, J. W. (2021). Extracting the dynamics of behavior in sensory decision-making experiments. *Neuron*, *109*(4), 597-610.e6. https://doi.org/10.1016/j.neuron.2020.12.004

Sayalı, C., Heling, E., & Cools, R. (2023). Learning progress mediates the link between cognitive effort and task engagement. *Cognition*, *236*, 105418. https://doi.org/10.1016/j.cognition.2023.105418

Schnabel, U. H., Van Der Bijl, T., Roelfsema, P. R., & Lorteije, J. A. M. (2021). A Direct Comparison of Spatial Attention and Stimulus–Response Compatibility between Mice and Humans. *Journal of Cognitive Neuroscience*, *33*(5), 771–783. https://doi.org/10.1162/jocn_a_01681

Shadlen, M. N., & Kiani, R. (2013). Decision Making as a Window on Cognition. *Neuron*, *80*(3), 791–806. https://doi.org/10.1016/j.neuron.2013.10.047

Sharpe, B. T., Trotter, M. G., & Hale, B. J. (2025). Sustaining student concentration: The effectiveness of micro-breaks in a classroom setting. *Frontiers in Psychology*, *16*, 1589411. https://doi.org/10.3389/fpsyg.2025.1589411

Smallwood, J., Brown, K. S., Tipper, C., Giesbrecht, B., Franklin, M. S., Mrazek, M. D., Carlson, J. M., & Schooler, J. W. (2011). Pupillometric Evidence for the Decoupling of Attention from Perceptual Input during Offline Thought. *PLoS ONE*, *6*(3), e18298. https://doi.org/10.1371/journal.pone.0018298

Sörensen, L. K. A., Bohté, S. M., Slagter, H. A., & Scholte, H. S. (2022). Arousal state affects perceptual decision-making by modulating hierarchical sensory processing in a large-scale visual system model. *PLOS Computational Biology*, *18*(4), e1009976. https://doi.org/10.1371/journal.pcbi.1009976

Sterzer, P. (2016). Moving forward in perceptual decision making. *Proceedings of the National Academy of Sciences*, *113*(21), 5771–5773. https://doi.org/10.1073/pnas.1605619113

Summerfield, C., & Blangero, A. (2017). Perceptual Decision-Making. In *Decision Neuroscience* (pp. 149–162). Elsevier. https://doi.org/10.1016/B978-0-12-805308-9.00012-9

The International Brain Laboratory, Aguillon-Rodriguez, V., Angelaki, D., Bayer, H., Bonacchi, N., Carandini, M., Cazettes, F., Chapuis, G., Churchland, A. K., Dan, Y., Dewitt, E., Faulkner, M., Forrest, H., Haetzel, L., Häusser, M., Hofer, S. B., Hu, F., Khanal, A., Krasniak, C., … Zador, A. M. (2021). Standardized and reproducible measurement of decision-making in mice. *eLife*, *10*, e63711. https://doi.org/10.7554/eLife.63711

Tsuda, B., Pate, S. C., Tye, K. M., Siegelmann, H. T., & Sejnowski, T. J. (2026). Neuromodulators Generate Multiple Context-Relevant Behaviors in Recurrent Neural Networks. *Neural Computation*, *38*(3), 292–327. https://doi.org/10.1162/NECO.a.1489

Unsworth, N., Redick, T. S., Lakey, C. E., & Young, D. L. (2010). Lapses in sustained attention and their relation to executive control and fluid abilities: An individual differences investigation. *Intelligence*, *38*(1), 111–122. https://doi.org/10.1016/j.intell.2009.08.002

Unsworth, N., & Robison, M. K. (2016). Pupillary correlates of lapses of sustained attention. *Cognitive, Affective, & Behavioral Neuroscience*, *16*(4), 601–615. https://doi.org/10.3758/s13415-016-0417-4

Urai, A. E. (2026). Structure uncovered: Understanding temporal variability in perceptual decision-making. *Trends in Cognitive Sciences*, *30*(1), 54–65. https://doi.org/10.1016/j.tics.2025.06.003

van den Brink, R. L., Murphy, P. R., & Nieuwenhuis, S. (2017). *Data from: Pupil diameter tracks lapses of attention* (Version 1, p. 1179880454 bytes) [Dataset]. Dryad. https://doi.org/10.5061/DRYAD.MP332

van den Brink, Ruud 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

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of 1/fa noise in human cognition. *Psychonomic Bulletin & Review*, *11*(4). https://doi.org/10.3758/BF03196615

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance Requires Hard Mental Work and Is Stressful. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 433–441. https://doi.org/10.1518/001872008X312152

Weiss, E., Liu, Y., & Wang, Q. (2026). The Contribution of the Locus Ceruleus–Norepinephrine System to the Coupling between Pupil-Linked Arousal and Cortical State. *The Journal of Neuroscience*, *46*(3), e0898252025. https://doi.org/10.1523/JNEUROSCI.0898-25.2025

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313. https://doi.org/10.3758/BF03194544

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, *18*(5), 459–482. https://doi.org/10.1002/cne.920180503

Appendix A

Data Inclusion Criteria

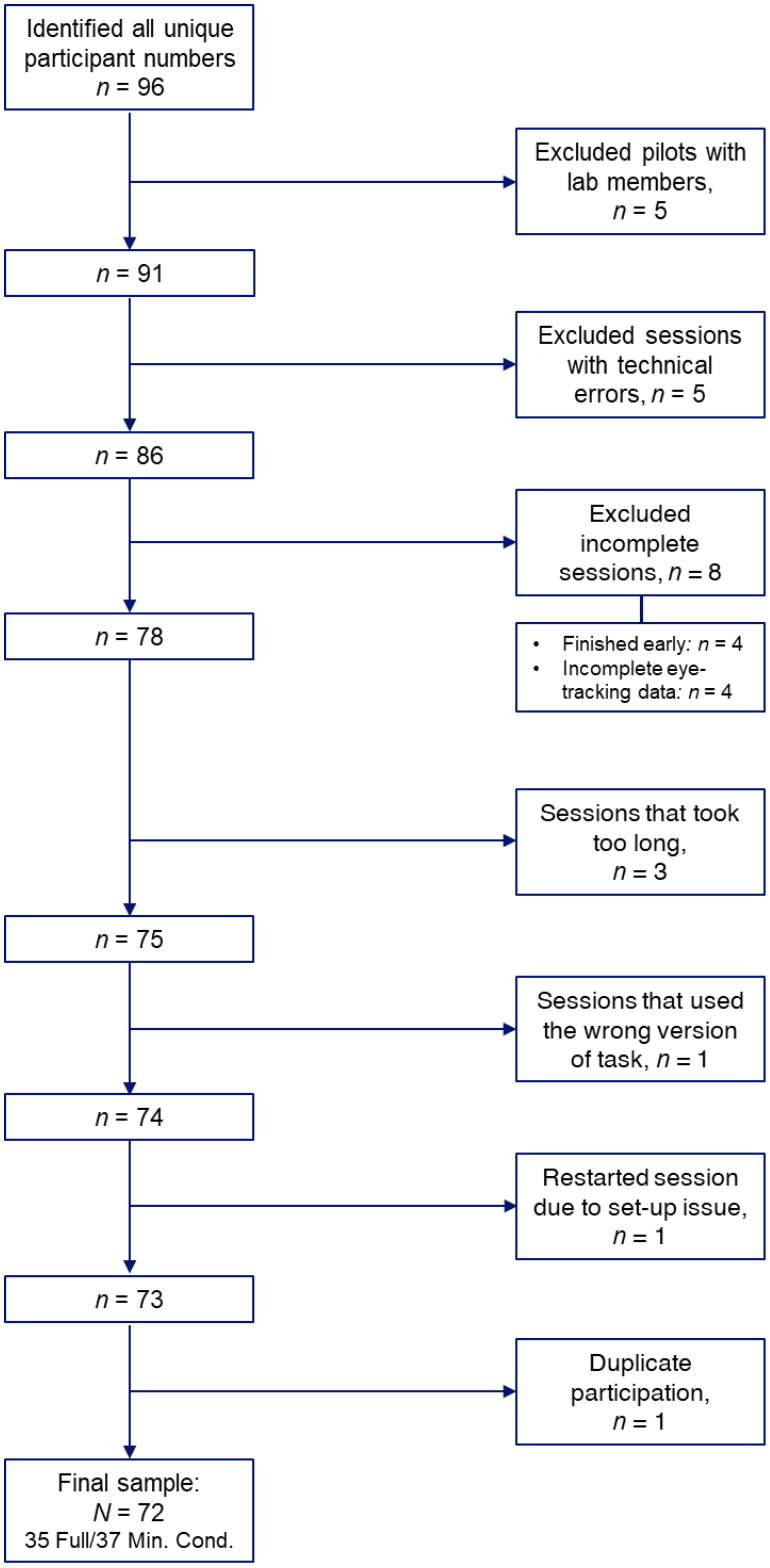


Figure B1. Data inclusion criteria based on post-hoc quality control of the collected data.

Appendix B
Instructions for the Experimental Conditions

Minimal Instructions

Thank you! We will now move on to the experiment. You will take part in a computer game. However, you will not receive any instructions: the goal is to figure out the rules of the game by yourself. To interact with the game, you will only need to move the mouse; clicking does not do anything. During the game, you will see a red cross in the middle of the screen. Please try to keep your eyes fixated on this cross throughout the experiment. When you are ready, click 'Start experiment'.

Full Instructions

Thank you! We will now walk you through the steps of this decision-making experiment. You will see some written instructions, as well as two example images of what the stimuli look like. You will then practice the task in a few example trials. The experiment consists of many trials. In each trial, you will see a red fixation point in the middle of the screen. Please keep your eyes focused on this fixation point throughout the experiment. At the beginning of each trial you will hear a sharp beep. You will then see two target images appear on each side of the screen. On each trial, you must decide which of the two target images has the higher contrast (i.e., appears darker). To select the target with the higher contrast, simply move the mouse until the chosen target reaches the center of the screen. For example, in the image below [Figure C1], the right target has the higher contrast, so you would move your mouse to the left. In this example image [Figure C2], the left target has the higher contrast, so you would move your mouse to the right. Simply moving the mouse left or right is enough, you don't need to click anything. You will hear a high beep if you answer correctly. If you answer incorrectly, you will hear a buzzing noise instead. If you take too long to answer, you will hear a low beep. Please try to be as accurate and fast as possible, and try not to let the trial time-out without a response. Click 'Continue' to practice a few example trials. That is the end of the practice trials. Here is a last reminder of the instructions: In each trial, you will hear a beep and two targets will appear. You must decide which target has the higher contrast. To respond, move your mouse left or right to bring the chosen target into the middle of the screen. You don't need to click, simply drag the mouse. If you are correct, you will hear a high beep; if you are incorrect, you will hear a buzzing noise. Please keep your eyes fixated on the red cross in the middle throughout the experiment. If you have any questions, feel free to ask out loud, the experimenter can hear you and will help you out. You will now start the real experiment. Good luck!

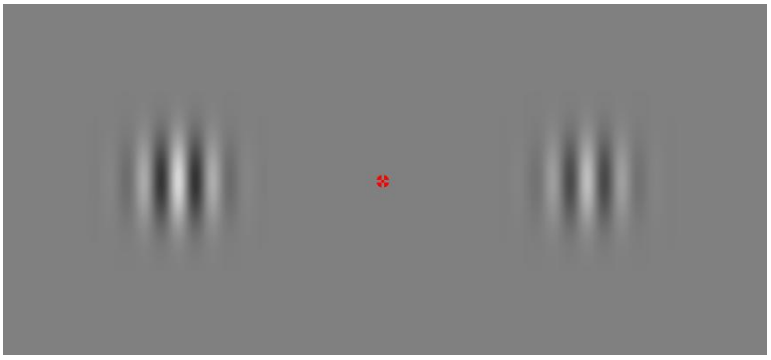


Figure B1. An example image where the target contrast is on the right.

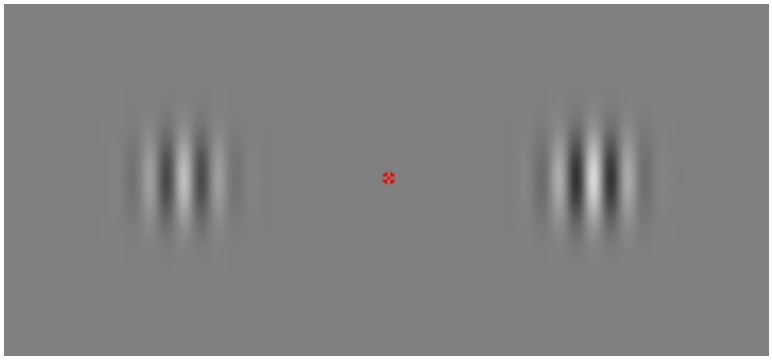


Figure B2. An example image where the target contrast is on the left.

Appendix C

Table of Demographics Characteristics

Table C1

Distribution of Gender and Handedness per Condition

Condition	Minimal instructions		Full instructions	
	n	*%*	*n*	*%*
Gender				
Female	31	84	29	83
Male	5	14	0	0
Non-binary	1	2	6	17
Handedness				
Right	33	89	30	86
Left	4	11	5	14

Note. $N = 72$ ($n = 37$ in minimal instructions condition, $n = 35$ in full instructions condition).

Appendix D

A Schematic Overview of the Sliding-Window Approach

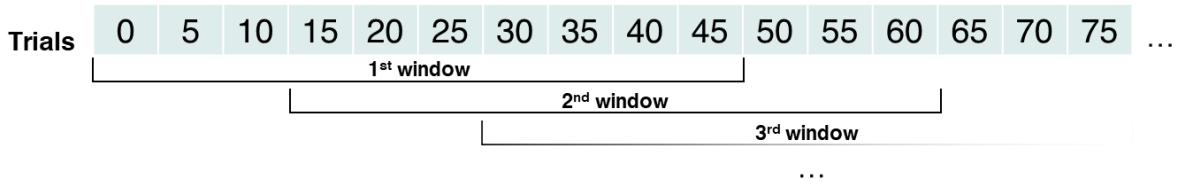


Figure D1. A sliding-window approach that was applied to each participant's performance data. Specifically, a window of 50 trials was moved across the dataset in steps of 15 trials, resulting in 37 overlapping windows.