

Pupillometry and the vigilance decrement: Task-evoked but not baseline pupil measures reflect declining performance in visual vigilance tasks

Joel T. Martin^{1,2}  | Annalise H. Whittaker³ | Stephen J. Johnston¹ 

¹School of Human and Health Sciences, Department of Psychology, University of Swansea, Swansea, UK

²Department of Psychology, University of York, York, UK

³Platform Systems Division, Defence, Science and Technology Laboratory (Dstl), Porton Down, Salisbury, UK

Correspondence

Joel T. Martin, Department of Psychology, University of York, York YO10 5DD, UK.
Email: joel.martin@york.ac.uk

Funding information

Defence Science and Technology Laboratory, Grant/Award Number: DSTLX1000083208

Edited by: John Foxe

Abstract

Baseline and task-evoked pupil measures are known to reflect the activity of the nervous system's central arousal mechanisms. With the increasing availability, affordability and flexibility of video-based eye tracking hardware, these measures may one day find practical application in real-time biobehavioural monitoring systems to assess performance or fitness for duty in tasks requiring vigilant attention. But real-world vigilance tasks are predominantly visual in their nature and most research in this area has taken place in the auditory domain. Here, we explore the relationship between pupil size—both baseline and task-evoked—and behavioural performance measures in two novel vigilance tasks requiring visual target detection: (1) a traditional vigilance task involving prolonged, continuous and uninterrupted performance ($n = 28$) and (2) a psychomotor vigilance task ($n = 25$). In both tasks, behavioural performance and task-evoked pupil responses declined as time spent on task increased, corroborating previous reports in the literature of a vigilance decrement with a corresponding reduction in task-evoked pupil measures. Also in line with previous findings, baseline pupil size did not show a consistent relationship with performance measures. Our data offer novel insights into the complex interplay of brain systems involved in vigilant attention and question the validity of the assumption that baseline (prestimulus) pupil size and task-evoked (poststimulus) pupil measures reflect the tonic and phasic firing modes of the locus coeruleus.

KEYWORDS

locus coeruleus, psychomotor vigilance, pupillometry, sustained attention, vigilance

Abbreviations: ANOVA, analysis of variance; *c*, criterion/response bias; *d'*, *d*-prime/perceptual sensitivity; EEG, electroencephalography; fMRI, functional magnetic resonance imaging; ISI, interstimulus interval; LC, locus coeruleus; NA, noradrenalin; PVT, psychomotor vigilance task; RT, reaction time.

1 | INTRODUCTION

The term *vigilance* has received varied usage in scientific research, but broadly speaking, it refers to an organism's ability to sustain its attention over prolonged periods of

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *European Journal of Neuroscience* published by Federation of European Neuroscience Societies and John Wiley & Sons Ltd.

time (Kahneman & Treisman, 1984; Parasuraman et al., 1998; Parasuraman & Davies, 1982; Warm et al., 2008; Warm & Jerison, 1984). Although there is a long history of research into performance and continuous work tasks (see Bills, 1943; Hockey, 2013 for reviews), Mackworth (1948, 1950) is frequently credited for the first systematic studies of vigilance and the discovery that human detection performance on a monotonous watch keeping task, under conditions similar to those experienced by radar and sonar operators, declines as time spent on task increases. This so-called *vigilance decrement* became the target of numerous research efforts in human factors and experimental psychology that sought to understand how factors specific to the task, the individual performing it, and the environment in which it is performed, all contribute to failures of vigilant attention (Frankmann & Adams, 1962; Mackie, 1987; Wiener, 1987). Signal detection theory (Green & Swets, 1974) has played a central role in the psychophysical analysis of vigilance studies, with detection performance being characterized frequently on the basis of the number of hits, misses, false alarms, correct rejections and the derived measures of sensitivity and criterion (e.g., Mackworth, 1970; Parasuraman & Davies, 1976). Detection latency also features in analyses of vigilance task performance (e.g., Basner & Dinges, 2011; Broadbent, 1958; Buck, 1966), and biometric technologies such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) continue to shape our understanding of the neurophysiological mechanisms of vigilant attention (for review, see Fortenbaugh et al., 2017; Langner & Eickhoff, 2013; Oken et al., 2006).

Another biometric technique that has been successfully applied to the study of vigilant attention is cognitive pupillometry, the measurement of the size and reactivity of the eyes' pupils following exposure to psychologically relevant stimuli. The pupils respond primarily to light, but when light levels are held constant, fluctuations in pupil size offer a window of insight into the brain's central arousal systems (Joshi & Gold, 2020; Kahneman, 1973; Laeng et al., 2012). Specifically, nonluminance-mediated pupil size changes are known to reflect the moment-to-moment activity of the locus coeruleus noradrenalin (LC-NA) system (Joshi et al., 2016; Rajkowski et al., 1993), which has a central role in the modulation of arousal and alertness (Berridge, 2008; Berridge et al., 2012; Berridge & Waterhouse, 2003) and in maintaining optimal levels of vigilance and performance (Petersen & Posner, 2012; Posner & Petersen, 1990). Extensive single-cell recording studies in behaving rodents and monkeys show that the noradrenergic neurons of the LC exhibit *phasic* and *tonic* modes of activation and that these distinct modes

correspond to different behavioural states (Aston-Jones et al., 1999; Aston-Jones & Cohen, 2005). The phasic mode is characterized by short bursts of activation in response to task-relevant stimuli and supports task engagement and exploitation of the environment, whereas the tonic mode is characterized by a sustained increase in baseline activation in response to diminishing task utility and supports disengagement from the current task and exploration of the environment (Aston-Jones & Cohen, 2005).

The functional role of the LC-NA system and its association with the pupil has led many to assume that baseline (i.e., prestimulus) and task-evoked measurements of pupil size may correspond to the tonic and phasic firing modes of the LC and that these measures may reflect changes in vigilant attention over time. Indeed, changes in pupillometric response associated with changes in vigilant attention have been noted previously. Beatty (1982) asked participants to monitor a string of tones presented at 3.2 s intervals continuously for 48 min for the occurrence of target tones, which were slightly attenuated in volume. Approximately 12 targets were presented at random intervals in every 5 min period, with 108 targets being presented across the whole task. As is common in vigilance tasks, although the task itself was conducted as one continuous 48-min procedure, the data were subdivided into several periods of watch to determine time-related differences in performance. Detection accuracy decreased in accordance with time spent on task, replicating the classic finding of a vigilance decrement described by Mackworth (1948, 1950). Task-evoked pupillary responses to target stimuli mirrored these results, decreasing in amplitude across each third of the test; but baseline (prestimulus) measurements of pupillary activity (obtained prior to each target stimulus) showed little change. In more recent work, we even see hints that pupil measures may serve to predict performance on a moment-to-moment basis. For example, Kristjansson et al. (2009) reported significant differences in pupil size and dilation rate for the fastest and slowest detection responses in a psychomotor vigilance task (PVT), suggesting that the pupil measures may provide sufficient reliable information to index alertness in real time. But, despite the promising narratives of Kristjansson et al. and a handful of other studies (e.g., Unsworth & Robison, 2016; van den Brink et al., 2016), the nature of the relationship between pupil and performance measures remains unclear.

Since Beatty (1982), many studies have found performance decrements in long and demanding tasks that coincided with reduced task-evoked pupil responses (e.g., Hopstaken et al., 2015a, 2015b; Murphy et al., 2016; Unsworth & Robison, 2016), but as noted by van den

Brink et al. (2016), the literature is conflicted on the relationship between task performance and baseline pupil size. In some experiments, moments of off-task thought or poor task performance were associated with larger pupils at baseline (Franklin et al., 2013; Gilzenrat et al., 2010; Smallwood et al., 2011, 2012; Unsworth & Robison, 2016), whereas in other experiments, poor task performance was associated with smaller pupils at baseline (Grandchamp et al., 2014; Hopstaken et al., 2015b; Kristjansson et al., 2009; Mittner et al., 2014; van Orden et al., 2000) or occurred following a gradual decrease in baseline pupil size (Grandchamp et al., 2014; Massar et al., 2016; McIntire et al., 2014; Murphy et al., 2011). Poor task performance was also found to occur with both relatively large and small baseline pupil size within experiments (Murphy et al., 2011; Smallwood et al., 2012; Unsworth & Robison, 2016; van den Brink et al., 2016), with one experiment reporting an increase in baseline pupil diameter as a function of time-on-task in a 37-min auditory vigilance task without breaks (Murphy et al., 2011).

Such discrepant findings on the relationship between baseline pupil size and task performance likely reflect the interplay of various methodological factors. Among the studies cited in the previous paragraph, there is considerable variability in how performance was measured, with some focusing primarily on reaction time (RT) measures, such as mean RTs (e.g., Smallwood et al., 2011, 2012), fraction of the slowest or fastest RTs (e.g., Unsworth & Robison, 2016; van den Brink et al., 2016) or RT variability (e.g., Murphy et al., 2011); and others focusing more on perceptual sensitivity (i.e., d' : Beatty, 1982; Hopstaken et al., 2015a, 2015b) or self-reported measures of task engagement (e.g., Franklin et al., 2013; Grandchamp et al., 2014; Mittner et al., 2014). Task demands also vary considerably across experiments, with some requiring only simple target detection (e.g., Massar et al., 2016) and others requiring *simultaneous* (Beatty, 1982; Gilzenrat et al., 2010; Murphy et al., 2011; van den Brink et al., 2016) or *successive* (Hopstaken et al., 2015a, 2015b, 2016; Smallwood et al., 2011, 2012) discrimination.¹ Further, some tasks called for prolonged continuous monitoring (Beatty, 1982; Murphy et al., 2011),

whereas others entailed intermittent breaks from the primary task (e.g., Hopstaken et al., 2015b; Smallwood et al., 2004; Unsworth & Robison, 2016), which even when very short have the potential to improve performance by temporarily boosting motivation (Ariga & Lleras, 2011; Ralph et al., 2016; Ross et al., 2014). Finally, the stimuli varied substantially, and some may have had undesirable behavioural or pupillometric consequences. For example, the 'running counter' stimulus used in Massar et al.'s (2016) PVT provides feedback that could enable participants to detect declines in their performance and adopt compensatory strategies (Thorne et al., 2005); and for studies using visual stimuli (e.g., Smallwood et al., 2011, 2012; van den Brink et al., 2016), differences in visual attributes such as luminance, colour and contrast may have contributed to pupillometric and behavioural variance (Barbur et al., 1992; Goldwater, 1972).

The increasing availability, affordability and flexibility of video-based eye tracking hardware means that pupils' predictive power for vigilant attention may one day find practical application in passive, real-time biobehavioural monitoring systems to assess performance or fitness for duty. Such systems would be of particular use in scenarios where traditional RT assessments cannot easily be administered. Presently, however, it remains unclear which pupil measures, if any, would be suitable for an application of this kind. Focusing on the issues raised above, here we examine the relationship between pupil measures—both baseline and task-evoked—and task performance in a novel implementation of two well-established vigilance task paradigms: (1) continuous, uninterrupted vigilance and (2) psychomotor vigilance.

2 | EXPERIMENT 1

Since Mackworth (1948, 1950), experimental vigilance tasks have generally aimed to simulate the conditions of real-world scenarios where monotonous repetitive tasks have become commonplace due to automation and industrial mechanization. Though vigilance tasks can vary in many ways, the defining characteristic is that observers must remain alert and respond to critical signals presented against a background of noncritical signals over prolonged, unbroken stretches of time—usually at least 30 min (Frankmann & Adams, 1962; Parasuraman & Davies, 1976). Key differences between tasks known to influence performance are the sensory modality of stimulus presentation (e.g., auditory and visual), the psychophysical dimensions used to define critical signals (e.g., brightness and loudness) and whether the detection of targets requires successive or

¹The distinction between successive and simultaneous discrimination tasks was first made by Parasuraman (1979). Successive tasks are absolute judgement tasks where observers must compare the current sensory input with a template in working memory in order to determine whether a particular stimulus is, or is not, a critical signal.

Simultaneous tasks on the other hand are comparative judgement tasks, where each stimulus contains all the information required to determine whether it is (or is not) a signal. Due to the involvement of working memory, successive tasks are thought to be more resource demanding than simultaneous tasks.

simultaneous discrimination (Parasuraman, 1979; Warm et al., 2008); but performance ultimately depends on complex interactions between factors relating to the task, the environment and the individual (Ballard, 1996). To date, most vigilance tasks conducted with pupillometry have presented stimuli in the auditory modality, probably to avoid methodological confounds associated with the effects of visual stimulation and optical distortion of raw pupil measurements inherent to video-based systems. But vigilance tasks in the real world are predominantly visual, and if pupil measures are to serve a useful purpose for tracking vigilant attention in real-world settings, they must be robust to the effects of visual stimuli.

In Experiment 1, we explored the relationship between pupil and performance measures (RT, accuracy, d' and c) in a canonical vigilance task with stimuli presented in the visual modality. To our knowledge, McIntire et al. (2014) is the only previous example of such a study, though the analysis was correlational and simply explored how average pupil size and percentage of hits showed a similar decline across four successive 10-min periods of continuous task performance. This is in contrast to auditory vigilance experiments (e.g., Beatty, 1982; Gilzenrat et al., 2010; Murphy et al., 2011), where there has been considerable focus on event-related pupil measurements. The present experiment therefore aimed to examine event-related pupil responses in a novel vigilance task with visual stimuli, whilst controlling appropriately for the effects of eye movements and luminance confounds.

Our task required participants to continuously monitor four centrally presented equiluminant visual stimuli for 30 min and to detect and respond to brief targets occurring with temporal and spatial uncertainty against a high background event rate. A relatively high number of targets—six per min—were used to ensure that a suitable amount of event-related data would be generated for the analysis (e.g., Mackie, 1987). We predicted that performance measures (e.g., RT, accuracy, d') across successive 10-min task blocks would betray a classic vigilance decrement as has been reported widely in the literature

(Frankmann & Adams, 1962; Mackie, 1987; Mackworth, 1948, 1950; Wiener, 1987). Second, on the basis of the most consistent findings from pupillometric studies of vigilance (e.g., Hopstaken et al., 2015a, 2015b; Murphy et al., 2016; Unsworth & Robison, 2016), we predicted that the magnitude of task-evoked pupil size changes would decrease, in line with performance measures, across the duration of the task. Considering the discrepant findings in the literature, we did not make specific predictions about baseline or 'tonic' pupil size for this experiment.

2.1 | Materials and methods

2.1.1 | Participants

Twenty-eight participants (23 females; age range 18–32 years, $M = 20.07$, $SD = 2.8$) completed the experiment voluntarily or in exchange for course credit. All participants were students at Swansea University reporting normal or corrected-to-normal acuity and colour vision. The experimental protocol was approved by the Ministry of Defence Research Ethics Committee and the Department of Psychology Ethics Committee at Swansea University. Written informed consent was obtained from each participant.

2.1.2 | Design

Task design reflected the core principles of classic experimental vigilance paradigms (Baddeley & Colquhoun, 1969; Mackworth, 1948, 1950; Parasuraman & Davies, 1976). Participants were asked to monitor four low-contrast gratings arranged squarely around a central fixation circle. The gratings rotated synchronously in a clockwise ticking motion at a rate of 120 ticks per minute (30° rotation per tick) and targets were defined as instances where one became briefly out of sync with the others (e.g., see Figure 1). The task lasted

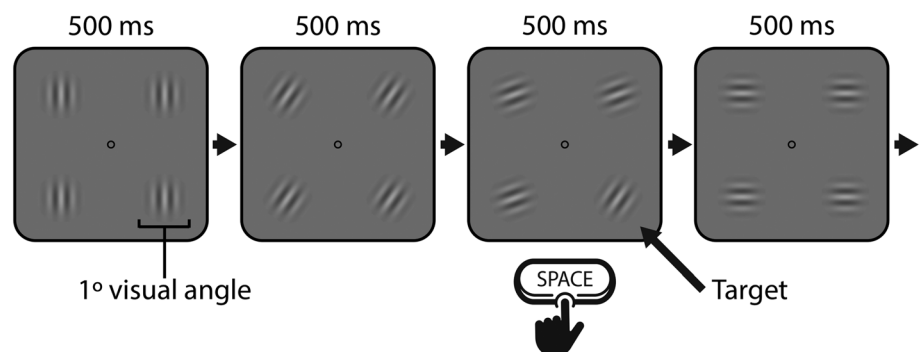


FIGURE 1 Schematic diagram of the stimuli and trial sequence for Experiment 1. Participants were asked to respond by pressing space every time a grating did not follow the standard sequence (6 times per min)

for 30 min, during which time continuous monitoring was required. Six targets were presented every minute (180 overall) at pseudorandom intervals, subject to the following constraints: (1) the time between targets was at least 6 s, and at most 30 s, (2) targets did not occur within 2 s of the beginning or end of the task. Targets occurred equally often at all of the four locations, although this was randomized across the whole experiment so that spatial uncertainty as to the location of the target would contribute to task difficulty (Broadbent, 1958; Mackie, 1987; Warm et al., 2008). All participants completed one trial of this experiment in a single testing session lasting approximately 40 min.

2.1.3 | Stimuli and apparatus

Four gratings enveloped within a cosine window (spatial frequency = 0.1, $SD = 12$, 39% contrast), each spanning $1^\circ \times 1^\circ$ of visual angle, were arranged in a square pattern around a central fixation circle on a grey background (Figure 1). The gratings were generated using an online tool² and saved in JPEG format. The task was administered on a 24-in. Iiyama monitor running at a resolution of $1,024 \times 768$ (1:1 aspect ratio) with a refresh rate of 144 Hz, and button responses were collected on a standard computer keyboard. The monitor and eye tracker were enclosed such that the only direct illumination came from the display screen and the participant could not see anything in their periphery. A viewing distance of 40 cm was maintained by a chin rest and forehead bar. Using a colorimeter (ColorCAL MKII, Cambridge Research Systems), the surface luminance of the grey background was recorded as 73.54 cd/m^2 , and the surrounding dark light of the unused portion of screen as 0.53 cd/m^2 . Pupil size and gaze data were recorded monocularly (left eye) with an EyeLink 1000 (SR Research, Mississauga, Ontario, Canada) system in tower mount configuration (32-mm lens) sampling at 250 Hz. According to the user manual, the system resolves pupil diameter to within 0.2 mm (SR Research Ltd., 2010). Eye level and camera position remained constant throughout the recording session for each participant. Stimulus presentation was managed with Experiment Builder (SR Research, Mississauga, Ontario, Canada).

2.1.4 | Procedure

On arrival at the lab, participants were told that for the next 30 min, they would be required to complete a

vigilance task that involved monitoring four circular patches rotating with a ticking motion at the centre of the screen. It was explained that, from time to time, one of the patches would briefly become out of phase with the others and that this was a target to which they had to respond. Participants were not given any further information about the frequency or temporal and spatial uncertainty of the targets. Once comfortable with the definition of a target, they were instructed that their task was to press the space bar every time they noticed such an event. Participants were forewarned that the task was monotonous but were asked to try and respond as quickly and accurately as possible. They were also instructed to maintain central fixation on the screen. A 5-point calibration and validation routine was performed prior to starting the experiment.

2.1.5 | Task performance and pupillometry

Task performance was assessed with RT, accuracy (i.e., percent hits and false alarms), and the signal detection theory measures sensitivity (d') and response bias (c). Hits, correct rejections, misses and false alarms were determined by an iterative algorithm that assigned button responses to stimulus events. For each button response, the RT in milliseconds to the last target was calculated. If the RT was greater than the minimum time between targets (6 s), or if no target had yet been presented, the button response was allocated to the nearest elapsed neutral event and labelled as a false alarm. All remaining button responses were then grouped together, and a permissible range for hits was determined as ± 2 median absolute deviations (Leys et al., 2013) from the group-level median RT. Accordingly, all button responses that occurred within 225–1,156 ms of targets were counted as hits, and those with RTs outside this range, as previous, were allocated to the nearest neutral event and counted as false alarms. The resulting distribution of RTs and the permissible hit range is illustrated in Figure 2. Finally, targets without a valid button response were counted as misses, and all remaining neutral events as correct rejections. This process of dealing with behavioural responses in sustained-attention tasks with high event rates is similar to that used by Esterman et al. (2016) and van den Brink et al. (2016). For the purposes of analysis, the complete experiment was decomposed into three 10-min periods of watch.

Individual pupil traces were extracted for all signal detection theory outcomes. For misses and correct rejections, the pupil data were time locked to the stimulus event, and baseline pupil size was defined as the average pupil size in a 500-ms window prior to the event. For hits

²<https://www.cogsci.nl/gabor-generator>

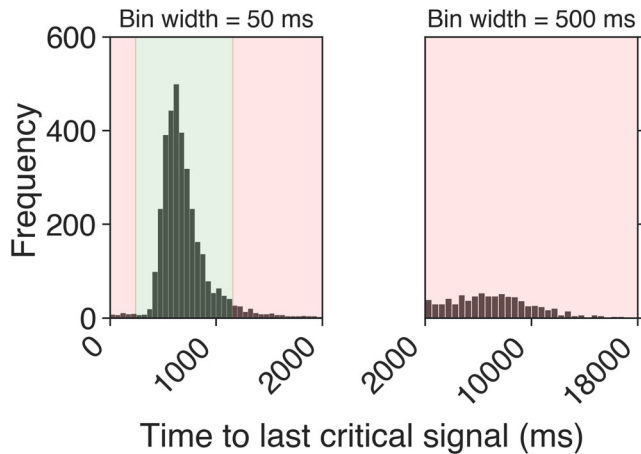


FIGURE 2 Distribution of reaction times (RTs) for Experiment 1 across all participants in 50-ms bins (left) and 500-ms bins (right). The green shaded area demarcates the permissible hit range (225–1,156 ms after the target), and the red shaded areas show the range where button responses were considered false alarms

and false alarms, the data were time locked to the corresponding button event, and the baseline period was offset by a further 500 ms (i.e., from $-1,000$ to -500 ms prior to the button response) to minimize contamination from fluctuations in pupil size associated with movement preparation and execution (Einhäuser et al., 2010; Hupé et al., 2009; Martin et al., 2020; Richer & Beatty, 1985).

Baseline and task-evoked pupil measures were also derived from the pupil traces to probe the effects of time-on-task. Baseline measures were defined as the average of the z -transform of pupil size (across the whole experiment) in the baseline periods, and task-evoked measures were defined as the average percent modulation in the peristimulus intervals (i.e., the portion of the pupil trace not including the baseline). We also averaged the z -transform of pupil size into 1-min bins for an overarching look at how pupil size varied across the whole task.

2.1.6 | Data processing and statistical analysis

Pupil data were processed and analysed using custom python scripts. Eye blinks were detected using the standard EyeLink parsing algorithm and reconstructed with linear interpolation prior to smoothing with a third-order Butterworth filter (4-Hz cut-off). Frequently, pre- and post-blink samples were noticeably part of the blink artefact, so we extended the blink endpoints by 100 ms in each direction. The average amount of data replaced by blink interpolation across all participants that were included in the analysis was 6.5%. Pupil data were then

down-sampled to 50 Hz, baseline corrected at the trial level with the subtractive procedure (Mathôt et al., 2018), and converted to units of percent signal change.

To examine the general pattern of pupil measures for each of the signal detection theory outcomes, stimulus-(misses and correct rejections) and button-locked (hits and false alarms) pupil traces from across the whole experiment were compared using two-tailed non-parametric permutation tests with cluster-based correction for the multiple comparisons problem (Maris & Oostenveld, 2007). This approach does not depend on theoretical assumptions about the data and reduces experimenter bias associated with choosing a time-period over which to compute summary statistics. For the permutation tests, t tests were used to compare two conditions (significance thresholds for test statistics determined theoretically from the appropriate degrees of freedom at $\alpha = .05$), and we follow the guidance of Sassenhagen and Draschkow (2019) for reporting and interpretation. To examine time-on-task effects, performance (RT, accuracy, d' , c) and scalar pupil measures were averaged within 10-min watch periods and analysed with repeated measures analysis of variance (ANOVA). Where Mauchly's W indicated that the assumption of sphericity was violated, p values were adjusted using the Greenhouse–Geisser correction. An alpha level of .05 was used for all statistical tests. The mean and standard deviation of horizontal ($M = 527$, $SD = 20$) and vertical ($M = 379$, $SD = 26$) gaze position for all samples included in the analysis indicate that participants maintained steady fixation at the centre of the screen throughout the task (Figure S1).

2.1.7 | Exclusions

Pupil data associated with stimulus and button events were discarded if the participant blinked during the baseline period or if more than 25% of the data across the epoch of interest were interpolated. Overall, this led to the discarding of pupil data for 17.89% of stimulus-locked (i.e., misses and correct rejections) epochs and 14.32% of button-locked (i.e., hits and false alarms) epochs. No participants were excluded from the analysis.

2.2 | Results

2.2.1 | Task performance

The average number of button responses made during the task was 151 ($SD = 51$). With respect to target events, the average percentage of hits and false alarms was

63.77% ($SD = 14.66\%$) and 1.09% ($SD = 1.43\%$), respectively. These data, together with the remaining behavioural data, are summarized in Figures 2 and 3. Figure 2 shows the RT distribution for the whole experiment including the permissible hit range (225–1,156-ms post-target), and Figure 3 shows accuracy (percentage of hits and false alarms), RT for hits, sensitivity (d') and response bias (c) as a function of watch period. The average RT for all hits was 666 ms ($SD = 156$ ms). Average sensitivity and response bias across the whole experiment were 2.93 ($SD = 0.67$) and 1.07 ($SD = 0.26$), respectively, indicating not only that perceptual sensitivity to targets was good but also that participants were generally biased to withhold responses to targets.

To examine the effects of time-on-task, one-factor (watch period) repeated measures ANOVAs were conducted for all performance measures. First, the percentage of hits and false alarms were analysed. There was a significant main effect of watch period on the percentage of hits, $F(2, 54) = 6.14$, $p = .004$, $\eta_p^2 = .19$, with Bonferroni-corrected t tests showing that participants attained a significantly higher percentage of hits in Watch Period 1 ($M = 69.2\%$, $SD = 14.0\%$) compared with Watch Period 3 ($M = 59.6\%$, $SD = 19.7\%$), $t(28) = 2.96$, standard error of the mean [SEM] = 3.21, $p = .019$. The difference between Watch Period 1 and Watch Period 2 ($M = 62.46\%$, $SD = 16.7\%$) was marginally significant, $t(28) = 2.37$, $SEM = 2.83$, $p = .076$, and the remaining comparison (2 vs. 3) was not significant ($p = .662$). The

effect of watch period on the percentage of false alarms was not significant ($p = .894$).

ANOVA on the RT data for hits revealed a significant main effect of watch period, $F(2, 54) = 16.51$, $p < .001$, $\eta_p^2 = .38$. Post hoc analysis with Bonferroni adjustment confirmed that RT was significantly faster in the first watch period ($M = 651$ ms, $SD = 79$ ms) compared with the second ($M = 687$ ms, $SD = 78$ ms), $t(28) = 4.95$, $SEM = 7.31$, $p < .001$, and the third ($M = 704$ ms, $SD = 89$ ms), $t(28) = 5.47$, $SEM = 9.79$, $p < .001$, but that there was no significant difference in RT between Watch Period 2 and Watch Period 3 ($p = .384$).

The same analysis was repeated for the signal detection measures sensitivity (d') and response bias (c). There was no significant main effect of watch period on sensitivity ($p = .193$), indicating that participants' ability to discriminate targets did not change throughout the task, but there was a significant main effect on response bias, $F(1.64, 44.31) = 7.38$, $p = .003$, $\eta_p^2 = .22$. Bonferroni-corrected post hoc t tests showed that response bias in Watch Period 1 ($M = 0.96$, $SD = 0.26$) was significantly lower than it was in Watch Period 2 ($M = 1.11$, $SD = 0.31$), $t(28) = 2.8$, $SEM = 0.05$, $p = .028$, and Watch Period 3 ($M = 1.15$, $SD = 0.33$), $t(28) = 3.11$, $SEM = 0.06$, $p = .013$, but that there was no significant difference between Watch Period 2 and Watch Period 3 ($p = .937$). This suggests that participants became more conservative as the task progressed and were therefore more reluctant to report that a target was present. As

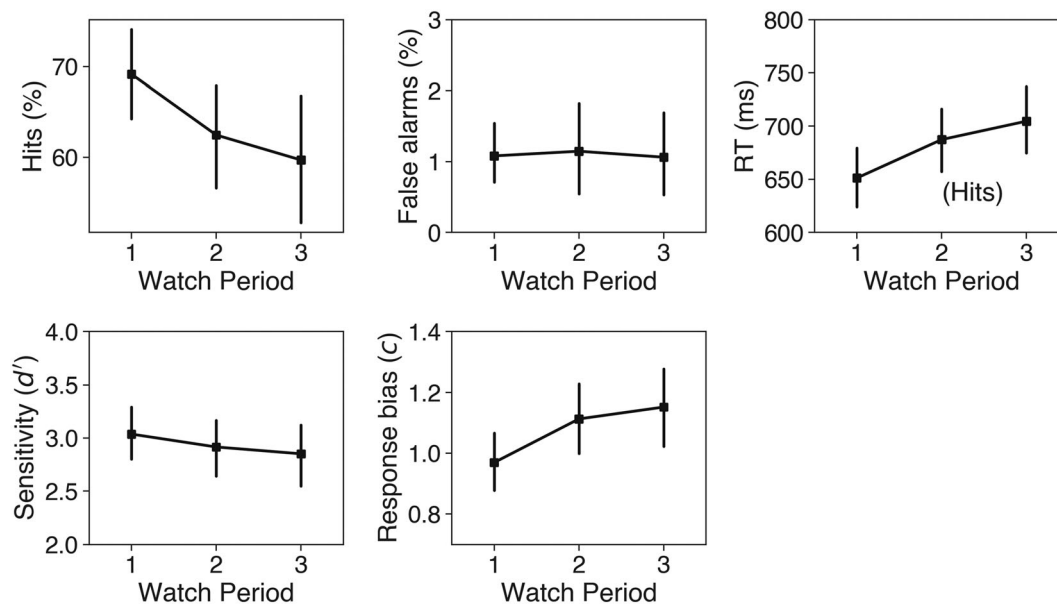


FIGURE 3 Task performance as a function of watch period in Experiment 1: Percentage of hits (top left) and false alarms (top middle), reaction times (RTs) for hits (top right), sensitivity (bottom left) and response bias (bottom middle). Error bars reflect 95% confidence intervals (bootstrapped, 1,000 iterations)

predicted, these performance data are consistent with the classic vigilance decrement.

2.2.2 | Pupil data

The z-transform of pupil data declined sharply in the first few minutes of the task and then increased steadily until the end. A cluster in the observed data extending from the fourth to the 12th minute differed significantly from the population mean (top panel, Figure 4). This time-on-task effect is well-noted in the literature for many different types of experiment (Fried et al., 2014; Hopstaken et al., 2015b, 2016; Massar et al., 2016; McIntire et al., 2014; Unsworth & Robison, 2016; van den Brink et al., 2016) and may reflect changes in overall arousal state or, more specifically, the transition from phasic to tonic modes of LC output (Joshi & Gold, 2020).

Event-related pupil data were time locked to button events for hits and false alarms and to stimulus events for misses and correct rejections. This was to ensure the

comparability of pupil data that were consistently affected by motor acts. The grand-average pupil traces for each of these behavioural outcomes are shown in Figure 4. Both button-locked outcomes (hits and false alarms) showed the usual pattern of pupil modulation associated with the preparation and execution of motor responses (e.g., see Einhäuser et al., 2010; Hupé et al., 2009; Martin et al., 2020; Richer & Beatty, 1985), with dilation beginning up to 500 ms before the motor act and peaking shortly afterwards. Permutation tests revealed significant modulation from baseline for hits and false alarms, as well as a significant difference between these two outcomes (lower right panel of Figure 4). The differences between the two traces can be summarized as follows. For hits, there was an average pupil modulation of 2.04% and a peak modulation of 5.62% with a latency of 400 ms from the button press, whereas for false alarms, these values were 3.55%, 7.07%, and 540 ms, respectively. For the stimulus-locked pupil measures (misses and correct rejections), only misses resulted in significant modulation from baseline in the poststimulus period, and there

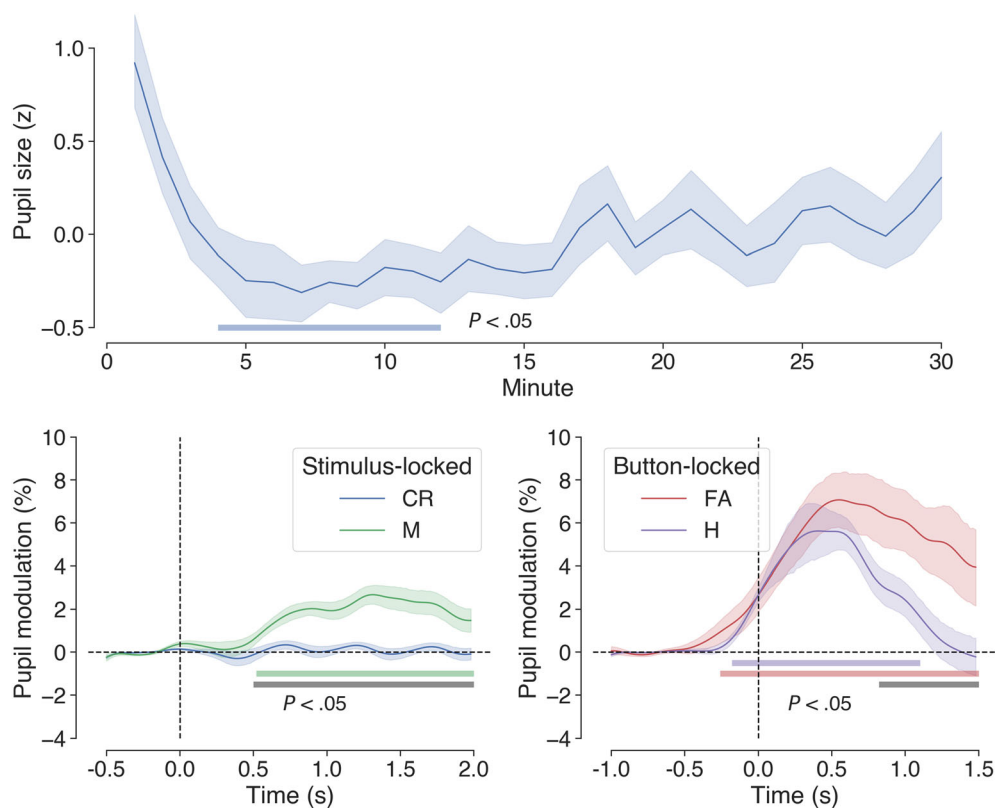


FIGURE 4 Grand-average pupil data in Experiment 1. Top panel shows the average z-transformed pupil data between subjects in 1-min bins across the whole experiment. A cluster extending from the fourth to the 12th minute differed significantly from the population norm (blue coloured bar). Bottom panels show stimulus- (left: correct rejections and misses) and button-locked (right: hits and false alarms) averages expressed as % change from baseline, with coloured horizontal bars indicating clusters of significant modulation from baseline and grey bars showing significant differences between traces (1,024 permutations, $p < .05$, cluster corrected for multiple comparisons). Shaded areas surrounding the pupil traces denote the standard error of the mean (SEM: Bootstrapped, 5,000 iterations)

was also a significant difference in pupil modulation between misses and correct rejections (bottom left panel of Figure 4). For misses, there was an average modulation of 1.25% and a peak modulation of 2.68% with a latency 1,320 ms. The time course for correct rejections resembled a flattened sine wave (average modulation of -0.05% , peak modulation of 0.34%) in phase with the onset of stimulus events. This periodic pattern is redolent of van den Brink et al.'s (2016) pupillometry data, which were observed in a task with a similar event-related design. We attribute this to task-correlated blinking and the blink-induced pupillary response (Knapen et al., 2016; Yoo et al., 2021).

After examining the stimulus- and button-locked pupil traces across the whole experiment, we probed the effects of time-on-task by analysing the scalar representations of pupil data with two-way (outcome \times watch period) repeated measures ANOVA. These data are displayed in Figure 5. For the stimulus-locked (i.e., misses and correct rejections) baseline pupil data, there were no significant effects of watch period ($F[2, 54] = 1.99$, $p = .146$) or outcome ($F[1, 27] = 0.12$, $p = .737$), and the outcome \times watch period interaction was not significant ($F[1.57, 42.39] = 0.37$, $p = .644$). For the stimulus-locked, task-evoked pupil modulations, there was a significant main effect of outcome, $F(1, 27) = 37.68$, $p < .001$, $\eta_p^2 = .58$, but the main effect of watch period ($F[2, 54] = 2.10$, $p = .132$) and the outcome \times watch period

interaction ($F[2, 54] = 2.01$, $p = .144$) were not significant. Simple main effects analysis showed that misses resulted in greater pupil modulation than correct rejections during Watch Period 1 ($F = 30.69$, $p < .001$) and Watch Period 2 ($F = 8.02$, $p = .009$), but not during Watch Period 3 ($F = 1.20$, $p = .283$).

For button-locked (i.e., hits and false alarms) baseline pupil data, the main effect of watch period was not significant, $F(1.58, 33.25) = 2.32$, $p = .124$, and the effect of outcome was marginally significant, $F(1, 21) = 4.03$, $p = .058$, with baseline pupil size being greater on average for false alarms compared with hits. The outcome \times watch period interaction did not significantly effect baseline pupil size, $F(2, 42) = 0.141$, $p = .869$. For the button-locked pupil modulations, the main effect of outcome was not significant, $F(1, 21) = 0.004$, $p = .953$, but there was a significant main effect of watch period $F(2, 42) = 8.86$, $p < .001$, $\eta_p^2 = .3$. Post hoc t tests revealed that the average % modulation for Watch Period 1 was greater than it was for Watch Period 2 ($MD = 2.32$, $t[22] = 2.95$, $p = .023$) or Watch Period 3 ($MD = 4.04$, $t[22] = 3.74$, $p = .004$), but that Watch Period 2 and Watch Period 3 did not differ significantly ($p = .297$). The outcome \times watch period interaction for button-locked pupil modulations was not significant ($F[1.52, 31.87] = 0.09$, $p = .861$).

Overall, these patterns in the pupil data are consistent with the prediction that the magnitude of task-evoked

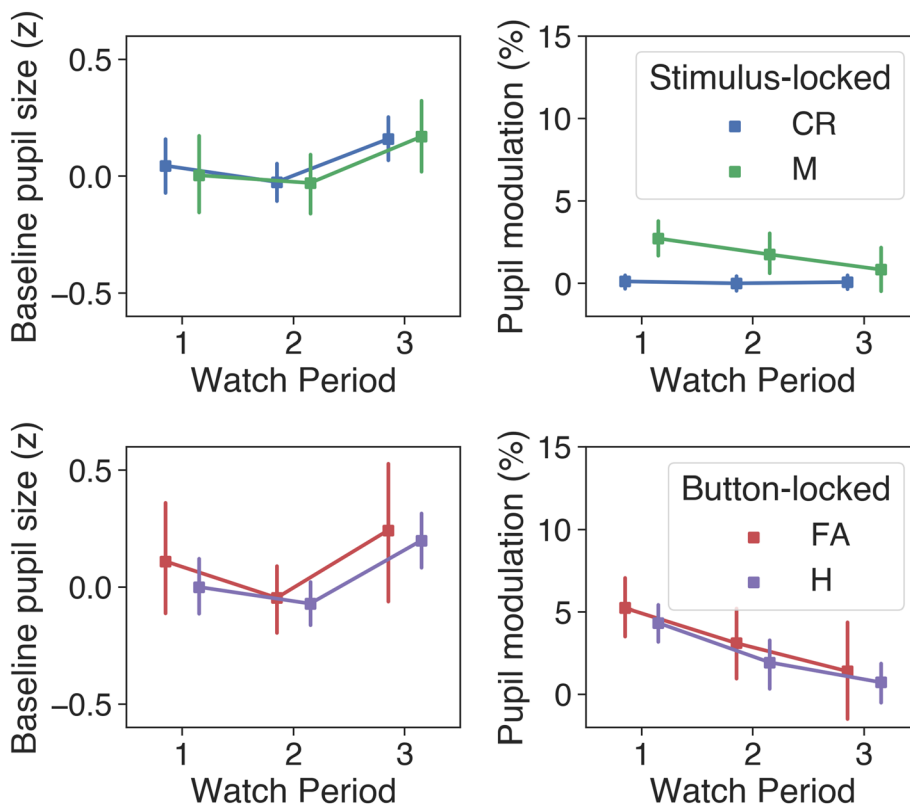


FIGURE 5 Stimulus- (M, misses; CR, correct rejections) and button-locked (H, hits; FA, false alarms) pupil averages across each watch period in Experiment 1. Error bars reflect 95% confidence intervals (bootstrapped, 1,000 iterations)

responses will mirror behavioural performance and decline as time-on-task increased.

2.2.3 | Correlational analyses

Across all button responses included in the analysis, RT did not significantly correlate with prestimulus baseline pupil size, $r(4102) = -.009$, $p = .559$, or task-evoked pupil size, $r(4102) = -.021$, $p = .073$. In line with previous literature (e.g., see de Gee et al., 2014), there was a significant negative correlation between baseline and task-evoked pupil size $r(4102) = -.273$, $p < .001$.

2.3 | Discussion

This experiment examined the relationship between task performance and pupil size in a prolonged, uninterrupted vigilance task with visually presented stimuli. Participants monitored four centrally located equiluminant gratings continuously for 30 min under the instruction to respond by pressing the space bar every time they detected a target. The task required successive discrimination (Parasuraman, 1979), but the high background event rate, the brief target duration and the temporal and spatial uncertainty of the targets added elements of difficulty (Broadbent, 1958; Mackie, 1987; Warm et al., 2008). In line with robust trends in the literature, we predicted that task performance and the magnitude of task-evoked pupillary responses would decrease as time-on-task increased.

Key behavioural measures were indicative of a classic vigilance decrement. Both the percentage of hits and the RT for hits changed across each successive 10-min watch period in a manner reflecting declining vigilance. These findings are consistent with well-established trends in the literature regarding the effects of time-on-task on detection performance under conditions of prolonged monitoring (e.g., Broadbent, 1953; Broadbent & Gregory, 1965; Buck, 1966; Mackworth, 1948, 1950; Parasuraman & Davies, 1976, 1982; Warm et al., 2008). The signal detection measures, sensitivity (d') and response bias (c), were calculated to gain further insight into the cause of the declining percentage of hits. Given that the nature of the task was in making trivially easy judgements about suprathreshold stimuli, it is not surprising that sensitivity remained at ceiling throughout. However, there was a conservative shift in response bias, suggesting that the decline in accuracy was linked to the participants' becoming less willing to report a detection, rather than a diminishing ability to discriminate targets from nontargets (Green & Swets, 1974). This is consistent with

previous reports that the vigilance decrement in tasks with high event rates is more closely related to changes in the strictness of the decision criterion over time, rather than perceptual sensitivity (e.g., Baddeley & Colquhoun, 1969; Broadbent, 1971; Colquhoun, 1961; Parasuraman & Davies, 1976).

Event-related pupil data were extracted for all signal detection outcomes to gain insight into the cognitive processing associated with these events. For misses and correct rejections, data were time locked to the onset of the relevant stimulus event (bottom left panel of Figure 4). Notably, misses resulted in reliably greater pupil dilation than correct rejections, a similar observation to that made by Beatty (1982) in his auditory vigilance experiment. As Beatty (1982) suggested, from a signal detection perspective, an enhanced pupillometric response to missed targets may reflect increased processing of sensory information for stimuli that fall close to the decision criterion. In this vein, the pupil dilation following missed targets may in part reflect subconscious processing of the target stimuli (Laeng et al., 2012). However, in the context of this experiment, such an interpretation must be tempered against the possibility that pupil modulation for misses was linked to detection, decision and motor effects of neighbouring button presses falling just outside of the permissible hit range. Previous research indicates that the pupillometric effects associated with motor acts can begin to emerge up to 1,000 ms prior to the act itself (e.g., Einhäuser et al., 2010; Hupé et al., 2009; Martin et al., 2020; Richer & Beatty, 1985), which means that false alarms whose RT fell just outside the permissible hit range (upper bound of 1,156 ms) may have contributed to the pupil dilation for missed targets. For hits and false alarms, pupil data were time locked to the button response and showed typical patterns of modulation associated with motor preparation and execution. Pupil dilation was also significantly greater for false alarms compared with hits (lower right panel of Figure 4). As suggested by Murphy et al. (2011), whose findings were similar, this difference may reflect the cognitive effects of a self-regulatory performance monitoring process. However, the larger pupil dilation for false alarms could also be linked to the higher degree of uncertainty that accompanies these events compared with correct detections (Yu & Dayan, 2005).

To examine the effects of time-on-task on pupil dynamics, scalar values of baseline and task-evoked pupil size were calculated for all stimulus and button events in each third of the task (Figure 5). As with Beatty (1982), baseline pupil size for all outcomes was relatively unchanged across the duration of the task, suggesting that these measures do not offer reliable insight into the

central processes underpinning the vigilance decrement. In contrast to this, however, task-evoked pupil modulations for misses, hits and false alarms exhibited a marked decline in magnitude across each successive watch period. This pattern of change parallels the decline in vigilance indexed by the percentage of hits and RT for hits and is therefore consistent with the findings from the vigilance experiments of Beatty (1982) and Murphy et al. (2011), as well as various other pupillometric studies of tasks requiring sustained attention (e.g., Hopstaken et al., 2015a; McIntire et al., 2014; Unsworth & Robison, 2016).

Task-evoked pupillary responses have been linked to phasic activation of the LC-NA system by neurophysiological and behavioural studies in both human and non-human primates (e.g., Alnaes et al., 2014; Aston-Jones & Cohen, 2005; Beatty, 1982; de Gee et al., 2017; Einhäuser et al., 2008; Gilzenrat et al., 2010; Jepma & Nieuwenhuis, 2011; Joshi et al., 2016; Murphy et al., 2011, 2014; Phillips et al., 2000; Rajkowski et al., 1993; Thorne et al., 2005; Urai et al., 2017; Varazzani et al., 2015). Further, single-unit recording studies in animals show that phasic activation of the LC-NA system occurs typically in response to task-related events during periods of high performance (e.g., Aston-Jones et al., 1991, 1994; Rajkowski et al., 1994). For example, the study by Aston-Jones et al. (1994) revealed that noradrenergic neurons in monkey LC are phasically activated by infrequent target cues during a vigilance task and also that the amplitude of these phasic responses diminishes over time. With respect to the well-known functional association between task-evoked pupil responses and phasic LC activation (Laeng et al., 2012), the findings from these animal studies fit well with those from the present study.

From a theoretical standpoint, it is difficult to ascertain whether the findings of this study fit best with a resource depletion, mind wandering or resource control-failure account of the vigilance decrement (see Caggiano & Parasuraman, 2004; Smallwood & Schooler, 2006; Thomson et al., 2015). Key indicators would be whether participants experienced the task as being effortful and the extent to which they engaged in task-unrelated thought, but we did not obtain these data as it would have required the use of intermittent thought probes (e.g., Hopstaken et al., 2015a; Smallwood et al., 2004; Unsworth & Robison, 2016), which involve temporary disengagement and therefore undermine a key aspect of classical vigilance task design—the requirement for continuous monitoring (Parasuraman & Davies, 1976). Thought probes may also serve as ‘mini breaks’, which can disrupt task monotony and alleviate the vigilance decrement (Ariga & Lleras, 2011; Ralph

et al., 2016; Ross et al., 2014). In the absence of subjective reports, only the pupil data and the nature of the task can serve as a basis for inferring the cause of the vigilance decrement. First, the task itself was prolonged and monotonous, and participants were without a strong incentive to maintain high levels of vigilance—conditions that provide fertile grounds for mind wandering (Smallwood & Schooler, 2006). Second, the magnitude of task-evoked dilations declined across the course of the task, an effect that has previously been linked to disengagement (Hopstaken et al., 2015b) and mind-wandering (Smallwood et al., 2011). Taken together, this suggests that the vigilance decrement in the current experiment may have been linked primarily to mind-wandering, or a resource control-failure leading to thought intrusion, but further data would be required to confirm this.

In sum, the present study replicates the well-known vigilance decrement—the reduction in detection performance that takes place during conditions of prolonged and continuous monitoring. Mirroring this behavioural effect, task-evoked pupil responses declined across the duration of the task, but baseline pupil size was mostly unchanged, suggesting that the vigilance decrement may have been linked to gradual disengagement of attention as the task progressed, rather than a change in organismic arousal state.

3 | EXPERIMENT 2

Traditional experimental vigilance tasks aim to emulate the conditions of real-world operator settings, but another form of vigilance task—the PVT (Wilkinson & Houghton, 1982)—aims to quickly assess declines in vigilant attention associated with sleep loss, circadian factors and other environmental stressors (Basner et al., 2011; Basner & Dinges, 2011; Blatter et al., 2006; Caldwell et al., 2003; Dinges et al., 1997; Graw et al., 2004; van Dongen & Dinges, 2005). In a PVT, instead of responding to infrequent signals over a prolonged period, subjects must make speeded responses to more regular signals occurring at pseudorandom intervals over a short period of time, usually 10 min or less. Wilkinson and Houghton's (1982) original version of this task was administered on a small hand-held battery-powered device displaying a millisecond counter set to ‘000’. The subject held the device and quickly pressed a button every time the counter began to increment, which happened at intervals ranging between 1 and 10 s. Upon detection of a response, the timer froze for 1.5 s, and the RT was saved before the timer reset to ‘000’. A variety of performance metrics can be derived from the data produced by this task, but analysis commonly focuses on

mean and median RT, the fastest and slowest 10% of trials, and the proportion of ‘lapses’, which are usually defined as RTs greater than 500 ms (Basner & Dinges, 2011).

In Experiment 2, we sought to examine how pupil measures relate to PVT task performance but with a novel stimulus approach optimized for pupillometry. Most PVTs utilize the prototypical stimulus of a running millisecond timer that counts up from zero but as noted by Thorne et al. (2005), this may have undesirable consequences. From a behavioural and pupillometric perspective, the two most relevant points made by Thorne et al. are as follows. First, the intensity of the stimulus changes in a non-linear fashion as the running counter increases, which could be a source of increasing variance in both behavioural and pupil measures. Second, the counter provides feedback to participants whether feedback is desired or not, which may enable them to monitor their own performance during the session and increase attention or effort to compensate for a noticed decline—a boon that would typically not be available in real-world settings. To avoid these issues, Thorne et al. (2005) devised a version of the PVT with a luminance-based graphic stimulus comprising two alternating black and white circular annuli resembling a target or ‘bull’s eye’. This PVT was administered on a dedicated hand-held device and produced highly comparable results despite its different stimulus characteristics.

Because a luminance-based stimulus such as that used by Thorne et al. (2005) would trigger the pupillary light reflex, the RT-initiating stimulus opted for in the present PVT experiment was a change in the orientation of a low-contrast grating. Participants were instructed to monitor a grating at the centre of a screen and respond as quickly as possible by pressing space whenever it flipped on its side (i.e., when it rotated 90°). As in other PVTs, the vigilance element of the task was instantiated with time-on-task (~13 min) and interstimulus interval (ISI) (4–12 s) parameters. Due to the use of a lower intensity stimulus, we expected that RTs in the current PVT would be slower on average than for PVTs using a running counter stimulus. However, because our approach avoids the confounds of variable stimulus intensity and feedback, behavioural and pupil measures should more faithfully reflect changes in vigilant attention. On the basis of the general findings outlined in the introduction, we predicted that time-on-task, both within and between successive blocks of the PVT, would lead to declining performance and a decrease in pupil size. Also, following the findings of Kristjansson et al. (2009) and Unsworth and Robison (2016), who employed similar tasks, we expected that poorest performance would be associated with smaller pupils at baseline.

3.1 | Materials and methods

3.1.1 | Participants

Twenty-five participants (18 females; age range 18–36 years, $M = 22.96$, $SD = 4.65$) completed the experiment voluntarily or in exchange for course credit. All participants were students at Swansea University reporting normal or corrected-to-normal acuity and colour vision. The experimental protocol was approved by the Department of Psychology Ethics Committee at Swansea University and the Ministry of Defence Research Ethics Committee. Written informed consent was obtained from each participant.

3.1.2 | Design

Performance and pupil measures were analysed in a repeated measures design as a function of trial group (1, 2, 3, 4 and 5: sets of 18 contiguous trials within a PVT block) and block (1, 2 and 3: successive blocks of 90 trials)—two factors whose purpose was to capture within- and between-block effects of time-on-task. The ISI, defined as the period between the last button response and the next flip of the grating, varied between 4 and 12 s. This period included a fixed component of 2,000 ms and a random component varying between 2 and 10 s. The random component was constrained such that a third of the intervals would be short (2,000–4,666 ms), a third medium (4,666–7,333 ms) and a third long (7,333–10,000 ms). The grating was present at the centre of the screen throughout the task.

3.1.3 | Stimuli and apparatus

The stimulus was a grating enveloped within a cosine window (spatial frequency = 0.1, $SD = 12$, 39% contrast) at the centre of the screen, spanning $2^\circ \times 2^\circ$ of visual angle. It was generated using the same process as specified for Experiment 1. All other details relating to hardware and software were the same as for Experiment 1.

3.1.4 | Procedure

Participants completed three consecutive blocks of the PVT in a single testing session, taking a forced break of only 1 min between blocks. In each block, participants were instructed to monitor the ‘circular stimulus’ at the centre of the screen and to respond as quickly as possible by pressing the space bar whenever it flipped on its side,

thereby to reset it to its original position (see Figure 6). Each block lasted approximately 13 min ($M = 12.8$, $SD = 0.63$), with some small variability arising from differences in RT and the random element of the ISI. Continuous recordings of gaze position and pupil data were obtained for each trial, and RT was defined as the time in milliseconds between the flip of the grating and the subsequent button response. A 5-point calibration and validation routine was performed at the start of each block. The task was performed in a dimly lit room.

3.1.5 | Data processing and statistical analysis

Pupil data were processed using the same general approach as described for Experiment 1. The average amount of data replaced by blink interpolation across all participants included in the analysis was 6.82%. Segments of pupil data 2,500 ms in length were extracted for each trial, time locked to the button response (−1,000 to 1,500 ms). These data were expressed as % modulation from a baseline defined as the average pupil size in a 500-ms period prior to the RT-initiating stimulus event.

To assess task performance, we focused on 1/RT and lapse frequency, which are among the most sensitive measures of alertness in PVTs (Basner & Dinges, 2011). Lapses in PVTs are traditionally defined as RT greater than 500 ms, but due to our novel take on the task, we defined lapses as RT greater than two median absolute deviations (Leys et al., 2013) from each participant median, which resulted in an average lapse threshold of 585 ms ($SD = 109$) across participants. Our pupil measures of interest were baseline pupil size and the task-evoked pupil response. Baseline pupil size was defined as the average z -transform of pupil size in the baseline period, whereas the task-evoked pupil response was defined as the average percentage of pupil modulation

around the time of the button response (−500 to 1,500 ms). The data for each of these four variables were analysed separately using two-factor (trial group \times block) repeated measures ANOVA. Where Mauchly's W indicated that the assumption of sphericity was violated, p values were adjusted using the Greenhouse–Geisser correction. We conducted further analyses on pupil measures for the fastest and slowest 20% RTs to determine whether pupil size at baseline was indicative of faster or slower detection responses (Kristjansson et al., 2009; Unsworth & Robison, 2016) and more generally how the extremes of performance are reflected in the pupil data. The mean and standard deviation of horizontal ($M = 518$, $SD = 10$) and vertical ($M = 388$, $SD = 14$) gaze position for all samples included in the analysis indicate that participants maintained steady fixation at the centre of the screen throughout the task (Figure S2).

3.1.6 | Exclusions

Two participants were excluded from the analysis for yielding poor quality pupil data (both had over 50% interpolated data for baselines and over 60% interpolated data for task-evoked responses). The general pattern of results was the same both with and without the exclusion of these participants. For the pupil analyses, trials were excluded if there was a blink in the baseline or if more than 25% of data were interpolated across the whole epoch (28.06% of trials).

3.2 | Results

3.2.1 | Task performance

Average RT across all participants was 420 ms ($SD = 79$) for non-lapse trials and 960 ms ($SD = 1,463$) for lapse

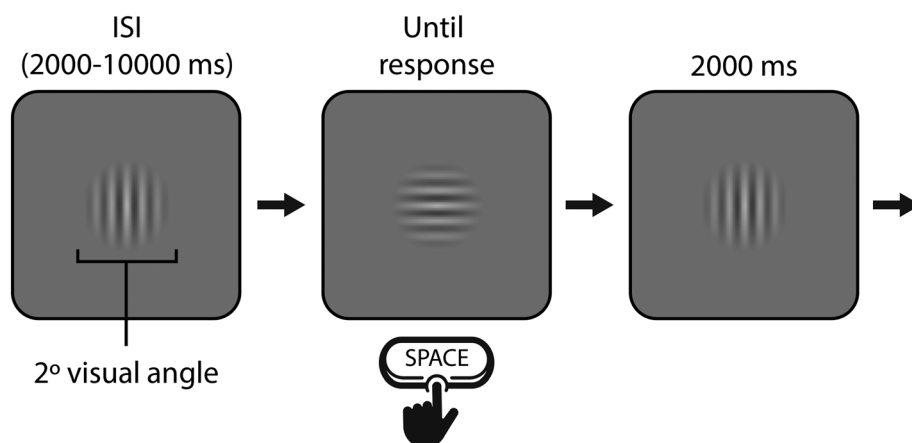


FIGURE 6 Schematic diagram of the stimulus and trial sequence for Experiment 2. Participants monitored a grating and responded by pressing space every time it flipped on its side (every 4–12 s)

trials. ANOVA on mean $1/RT$ (i.e., the reciprocal transform of RT) revealed a significant Trial Group \times Block interaction, $F(8, 176) = 3.16$, $p = .002$, $\eta_p^2 = .13$, the nature of which is illustrated in the top left panel of Figure 7. Simple effects showed that $1/RT$ decreased significantly across trial group in each block (Block 1: $F = 16.65$, $p < .001$; Block 2: $F = 2.89$, $p = .027$; Block 3: $F = 3.40$, $p < .012$), which is consistent with the prediction that performance would decline as time-on-task increased. Post hoc analysis with Bonferroni adjustment showed that, in Trial Groups 1–3, $1/RT$ was significantly greater in Block 1 compared with Blocks 2 and 3 (all $ps < .05$) and that $1/RT$ in Trial Group 4 was significantly greater for Block 1 compared with Block 3 ($p < .05$). No other comparisons were significant ($p > .05$). Therefore, as indexed by $1/RT$, performance was best overall in Block 1 compared with Blocks 2 and 3, but the magnitude of this effect decreased across trial groups.

Average lapse frequency across participants was 27.6 ($SD = 7.8$). ANOVA showed that the main effect of lapse frequency was significant for trial group, $F(4, 88) = 5.43$, $p < .001$, $\eta_p^2 = .20$, and for Block, $F(2, 44) = 16.74$, $p < .001$, $\eta_p^2 = .43$ (top right panel of Figure 7), but that the trial group \times block interaction was not significant ($F[8, 176]$, $p = .541$). Simple effects for trial group showed that the number of lapses increased significantly throughout Block 1 ($F = 3.16$, $p = .018$) and Block 3 ($F = 3.01$, $p = .022$), but not Block 2 ($F = 0.60$,

$p < .662$). Lapse frequency therefore followed the same general pattern as $1/RT$, and together these data support the prediction that performance would decline as time-on-task increased.

3.2.2 | Pupil data

Grand-average button-locked pupil traces for each block are shown in Figure 8. The pupil began to dilate slowly following the stimulus event and then rapidly after the button press. In the 1,500 ms following the button press, there was an average modulation of 5.22% and a peak latency of 880 ms. A conspicuous trough in the pupil traces after the button press coincides with a transient but marked increase in the percentage of interpolated data. This artefact resembles the blink-induced pupillary response (e.g., Knapen et al., 2016; Yoo et al., 2021) and is therefore indicative of task-correlated blinking (i.e., participants tended to blink after button presses). We did not correct this artefact with linear interpolation as it would involve altering too much data and excluding more trials.

ANOVA on the baseline pupil measures revealed a significant trial group \times block interaction, $F(3.81, 76.10) = 3.47$, $p = .013$, $\eta_p^2 = .15$, which is displayed in the bottom left panel of Figure 7. Simple effects analysis for trial group showed that baseline pupil size decreased

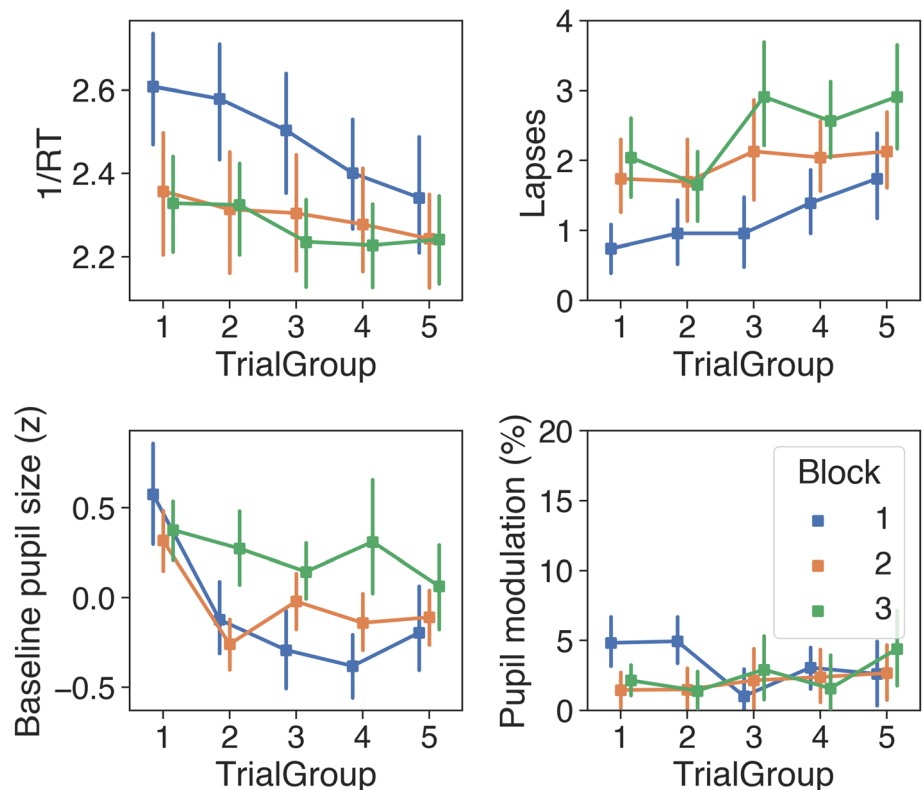


FIGURE 7 Performance (top row) and pupil (bottom row) measures across trial group and block in Experiment 2, with error bars showing 95% confidence intervals (bootstrapped, 1,000 iterations)

significantly across Block 1 ($F = 13.57$, $p < .001$) and Block 2 ($F = 7.02$, $p < .001$), but not Block 3 ($F = 0.89$, $p = .470$). Bonferroni-corrected post hoc t tests revealed that baseline pupil size was significantly greater in Trial Group 4 for Block 3 compared with Block 1 ($p = .007$), but no other between Block comparisons were significant ($p > .05$).

For measures of pupil modulation, there was a significant main effect of Block, $F(2, 42) = 5.93$, $p = .010$,

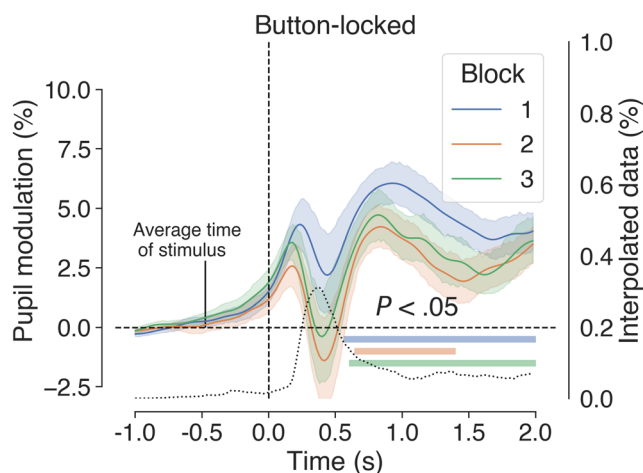


FIGURE 8 Average button-locked pupil traces for each block in Experiment 2. Black dotted trace shows the percentage of interpolated data, which indicates task-correlated blinking. Shaded areas surrounding the coloured traces show the standard error of the mean (SEM) (bootstrapped, 5,000 iterations) and coloured horizontal bars denote clusters of significant modulation from baseline, as revealed by non-parametric permutation tests (1,024 permutations, $p < .05$, cluster corrected for multiple comparisons)

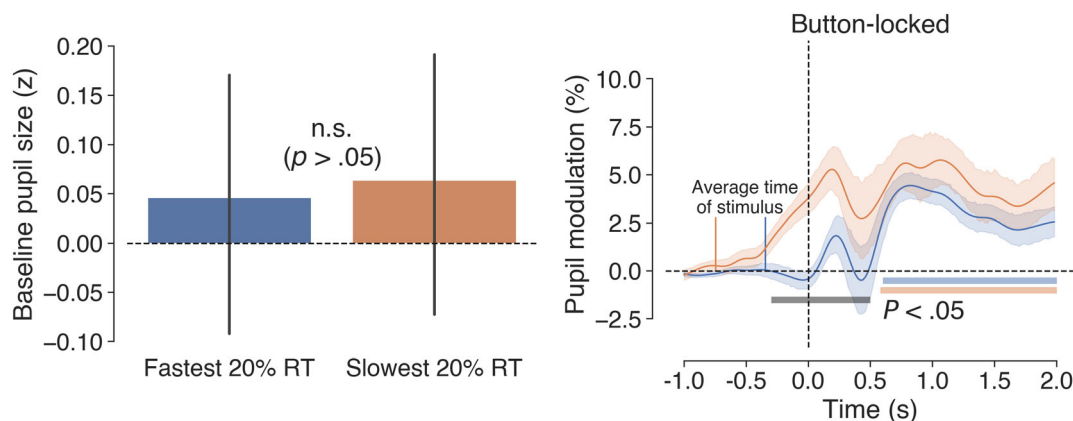


FIGURE 9 Baseline and button-locked pupil measures for the fastest ($M = 347$ ms, $SD = 43$ ms) and slowest ($M = 746$ ms, $SD = 43$ ms) 20% reaction times (RTs) in Experiment 2. Left panel shows mean prestimulus baseline pupil size with 95% confidence intervals (bootstrapped, 1,000 iterations), and the right panel shows pupil dilations time locked to button responses, with shaded areas surrounding the pupil traces denoting the standard error of the mean (SEM) (bootstrapped, 5,000 iterations). Coloured horizontal bars in the right-hand panel denote clusters of significant modulation from baseline for the respective traces (grey bar represents the difference between the traces), as revealed by non-parametric permutation tests (1,024 permutations, $p < .05$, cluster corrected for multiple comparisons)

$\eta_p^2 = .23$, but the effect of trial group ($F[4, 80] = 1.86$, $p = .125$) and the trial group \times block interaction ($F[4.83, 96.63] = 1.87$, $p = .109$) were not significant. Simple main effects showed that pupil modulation was greater in Block 1 for Trial Group 1 ($F = 8.97$, $p < .001$) and 2 ($F = 9.07$, $p < .001$), but not for Trial Groups 3 to 5 ($p > .05$).

3.2.3 | Correlational analyses

Across all trials included in the analysis, RT correlated significantly with task-evoked pupil size, $r(4466) = -.159$, $p < .001$, but not with baseline pupil size, $r(4466) = -.011$, $p = .459$. The classic negative correlation between baseline and task-evoked pupil size (e.g., see de Gee et al., 2014) was also present, $r(4466) = -.366$, $p < .001$.

3.2.4 | Fastest versus slowest RTs

To explore how the extremes of performance are reflected in the pupil data, we conducted further analysis on trials with the fastest ($M = 347$ ms, $SD = 43$ ms) and slowest ($M = 746$ ms, $SD = 43$ ms) 20% RTs. Figure 9 displays the pupillometry results for these extreme quintiles. Baseline pupil size did not differ significantly ($p > .05$, left panel of Figure 9), but there was a significant difference in the button-locked pupil traces ($p < .05$, cluster-corrected permutation test, right panel of Figure 9) marked by a cluster spanning the button event. This difference clearly pertained to the timing and

magnitude of pupil dilation. For the slowest RTs, dilation began prior to the button response, whereas for the faster RTs, dilation did not begin until afterwards. Further, the average modulation in the 2,000-ms post-button period was greater on average for the slowest ($M = 4.39\%$, $SD = 5.46\%$) than for the fastest RTs ($M = 2.45\%$, $SD = 3.23\%$). These data do not corroborate previous reports of pretrial baseline predicting performance (e.g., Kristjansson et al., 2009; Unsworth & Robison, 2016), but rather they suggest that, at least within the context of our experiment, the pattern of pupil dilation prior to a detection response may be the more relevant predictor. In this respect, our data are in line with recent PVT studies where the fastest RTs were associated with larger pupil dilations in the ISI (Unsworth et al., 2020; Unsworth & Robison, 2018).

3.3 | Discussion

This experiment sought insight into the relationship between pupil size and performance measures in a novel PVT. Participants monitored a low-contrast grating for a sudden 90° rotation and responded with a button press as quickly as possible after the event. We adopted an atypical stimulus approach to avoid confounds associated with the canonical running counter stimulus—namely, its variable intensity and the performance feedback that it provides—which could potentially contribute to variance in behavioural and pupillometric measures (Thorne et al., 2005). Participants completed three successive blocks of the task taking only a 1-min break in between, and changes in performance and pupil measures were explored both within and between blocks. We predicted that performance and pupil size would decrease as time-on-task increased and that worse performance would be associated with smaller pupils at baseline.

It is noteworthy that our novel stimulus approach led to longer RTs than are typically observed in PVTs that use the canonical running counter stimulus. In these tasks, average RT for subjectively alert participants is generally in the range of 200–300 ms (e.g., Basner et al., 2011; Blatter et al., 2006; Dorrian et al., 2007; Loh et al., 2004; Matsangas et al., 2016; McClelland et al., 2010; Wilkinson & Houghton, 1982), whereas in the current PVT, also with subjectively alert participants, average RT was 420 ms. We attribute this to differences in stimulus intensity. The running counter stimulus is dynamic and constantly changing, providing a constantly refreshed cue for the participant to respond, whereas a change in the orientation of a low-contrast grating is more subtle and discrete, and issues no refreshing cue to respond.

As predicted, the main performance measures exhibited typical time-on-task effects, with 1/RT decreasing and the number of lapses increasing as time-on-task increased. This general pattern was observed within and between each block of the PVT for both performance measures. For 1/RT, the biggest change was between the first block and the two subsequent blocks, with the difference being largest across the first three groups of trials. Lapse frequency increased gradually within each block and between successive blocks. These patterns in the performance data were statistically robust even without the state manipulations (e.g., time of day and sleep deprivation) and large number of repeated tests that are often integral to the design of mainstream PVT research (e.g., Basner et al., 2011; Blatter et al., 2006; Dorrian et al., 2003; Graw et al., 2004; Loh et al., 2004; Manousakis et al., 2017).

As regards the pupil data, the pattern of within-block declining baseline pupil size broadly reflected the decline in task performance, corroborating findings from previous PVT studies (Massar et al., 2016; Unsworth & Robison, 2016) as well as various other studies that examined pupil and performance measures in tasks requiring vigilance or sustained attention (van den Brink et al., 2016; Grandchamp et al., 2014; Hopstaken et al., 2015a; McIntire et al., 2014; van Orden et al., 2000). However, the relationship between baseline pupil size and task performance was not clear cut. Participants performed best and had the largest baseline pupil size at the beginning of each block, but the sharp drop in baseline pupil size that occurred between the first and second trial group did not have a commensurate drop in performance. Further, baseline pupil size was largest overall and showed the least variability across trial groups in Block 3, where performance was at its worst. In a similar fashion, the task-evoked pupil responses were largest at the beginning of Block 1, where performance was best, but were less consistent with respect to the performance data at other times. These patterns in the pupil data are in line with the general prediction that pupil size would decrease as time-on-task increased, but they run contrary to the prediction that worse performance would be reflected in smaller pupils at baseline.

Previous experiments offer conflicting evidence as to whether optimal task performance is associated with larger or smaller pupils at baseline (e.g., Kristjansson et al., 2009; Unsworth & Robison, 2016). To address this issue, we compared baseline and task-evoked pupil responses for the trials with the fastest and slowest 20% RTs. Although there was no significant difference in baseline pupil size between these two groups of trials, there was a clear difference between the observed pupil traces. For the faster RTs, dilation did not begin until

after the button response was made, whereas for the slower RTs, dilation was apparent around 500 ms before the button response and increased gradually until it peaked shortly afterwards. The finding of gradual dilation prior to an overt detection response, which is well documented in the literature, has been linked to cognitive factors associated with target recognition and decision making (e.g., see Einhäuser et al., 2010; Martin et al., 2020; Richer & Beatty, 1985). The reason we see this only for the slowest and not the fastest trials probably reflects the difference in RT and the fact that genuine cognitive effects on pupil size tend not to develop until at least 220 ms from the causal event (Mathôt et al., 2015, 2018). For the fastest trials, pupil modulation effects relating to target recognition and decision making were likely mixed in with the motor component.

The findings from the current experiment are generally consistent with previous studies showing time-on-task effects on performance and pupil size, but they do not align perfectly with a specific theory of vigilance. The understimulating and unrewarding nature of the task does however provide ripe conditions for mind-wandering, suggesting that this may have been partly responsible for the decline in performance. Previous studies have also reported larger pupils at baseline during periods of mind-wandering and poor task performance (e.g., Franklin et al., 2013; Smallwood et al., 2011, 2012; Unsworth & Robison, 2016), which is the pattern that was observed in Block 3 of the current experiment. Research also suggests that very short breaks can reduce mind-wandering and lead to performance improvements by temporarily boosting motivation (e.g., Ariga & Lleras, 2011; Ralph et al., 2016; Ross et al., 2014), which fits with the pattern of data in the current experiment, where participants' performance was restored to more optimal levels after taking a 1-min break in between each block.

We recognize that various factors relating to the individual state of the participants could have influenced the results of the present experiment. For example, performance in PVTs is affected by sleep pressure (Blatter et al., 2006), time of day and its interaction with circadian rhythms (Graw et al., 2004; van Dongen & Dinges, 2005), the consumption of stimulants such as caffeine (van Dongen et al., 2001), and individual differences in intrinsic alertness (Unsworth et al., 2020). The current experiment did not control for any of such factors, but this could easily be achieved in a subsequent study. For instance, circadian effects could be controlled for by excluding strong 'morning and evening types' (Horne et al., 1980) and by testing participants at the same times during the day, after they have reported having similar amounts of sleep. Alternatively, one could examine how performance and pupillometry vary with respect to

individual differences in a broad range of cognitive and self-reported personality factors (e.g., Unsworth et al., 2019, 2020).

Finally, we note that our novel take on the PVT limits the extent to which it can be directly compared with a more traditional PVT. The use of an alternative stimulus was desirable to avoid certain confounds, but the experiment also differed in terms of block length and ISI. In their general recommendations for the standardized design and analysis of PVTs, Basner and Dinges (2011) suggest using an ISI of 2–10 s and having a fixed block length of 10 min. Due to the way the current experiment was implemented, ISI was 4–12 s, and block length was variable ($M = 12.8$ min, $SD = 0.63$ min). Future experiments may wish to bring our approach closer to the task specifications set out by Basner and Dinges (2011), which would broaden the basis for comparison of experimental findings in the wider literature.

4 | GENERAL DISCUSSION

Recent pupillometric studies of vigilance and sustained attention suggest that measurements of pupil size could potentially be used in real-world settings to monitor performance, and perhaps even to predict and prevent errors associated with lapses of attention before they occur. But the literature in this area—especially regarding visual tasks—is sparse, and differences in methodology and task requirements have led to conflicting findings. The purpose of this study was to further explore the relationship between pupil size and performance measures in the context of well-established task frameworks from the vigilance literature.

The most consistent finding across both experiments regarding the relationship between pupil size and monitoring performance was that, in line with previous experimental findings (e.g., Beatty, 1982; Hopstaken et al., 2015b; Unsworth & Robison, 2016) and the predictions of established theory (Aston-Jones & Cohen, 2005), task-evoked pupil responses were generally more pronounced when performance was best. This trend was most consistent in Experiment 1, where the decline in detection performance was mirrored by a decline in the magnitude of task-evoked responses associated with hits, misses and false alarms. In Experiment 2, the relationship between task-evoked responses and performance measures was less consistent, although the largest responses did occur when performance was best (i.e., at the beginning of Block 1). In general, these findings suggest that changes in task-evoked pupil responses may serve as an accurate indication of general task engagement, with a decline in their magnitude over time reflecting cognitive

disengagement from the task and an increased likelihood of suboptimal performance.

Our baseline pupil measures did not show a consistent relationship with performance. In Experiment 1, baseline pupil size was mostly unchanged across three successive periods of watch, despite a marked decrement in performance. In Experiment 2, baseline pupil size showed an overall decline within each Block, although the slope became less pronounced with each successive Block. Interestingly, baseline pupil size was biggest overall at the beginning of each Block, where task performance was best, suggesting that it may be related to heightened arousal, alertness and focused attention. But, by this account, our baseline measures in the PVT reflect combinations of autonomic tone as well as task-related factors, which means that they are not serving uniquely as a window of insight into the ‘tonic’ mode of LC activation, as is often explicitly or implicitly assumed (see below). The lack of consistency in our baseline measures and their relationship with performance metrics is not unprecedented in light of the literature reviewed in the introduction, which indicates that the relationship is complex and in need of further characterization. One possibility raised by van den Brink et al. (2016) is that the effects of time-on-task on baseline pupil size obscure a more nuanced relationship with performance. In their gradual-onset performance task, after regressing out the effects of time-on-task from the baseline pupil data, the authors observed a quadratic relationship with performance, such that performance was optimal when baseline pupil size was at intermediate levels. This idea dovetails with the Yerkes–Dodson law (Yerkes & Dodson, 1908) of optimum arousal, whereby the relationship between task performance and arousal is described by an inverted-U function, such that poor performance is associated with both under- and over-arousal, and optimum performance occurs at a ‘sweet spot’ on the arousal curve.

We refrained from using the words ‘tonic’ and ‘phasic’ to describe our pupil measures because we are aware of numerous caveats to the assumption that baseline and task-evoked measures map neatly onto the different modes of LC output. Joshi and Gold (2020) discuss this issue in detail and emphasize that, in the context of LC activation, the terms ‘tonic’ and ‘phasic’ differentiate between distinct modes of activation, and not simply between baseline and transient activity (Aston-Jones & Cohen, 2005). This is complicated further by the substantial variability between publications in how ‘tonic’ and ‘phasic’ pupil measures are operationally defined. Further, the precise neural mechanisms of the relationship between pupil measures and LC activation are presently unclear, and it is possible that a third variable, as of yet

not understood, may account for the observed pupil-LC link (Costa & Rudebeck, 2016).

In conclusion, the results of our two vigilance experiments support the general notion that changes in task-evoked pupil measures can be used to gain insight into monitoring performance and organismic arousal state in long and demanding tasks where the emphasis is on aggregate effects over a series of trials. But there is clearly a need for further research to determine the practical feasibility of utilizing pupil size as a psychophysiological marker for attentional lapses in real-time monitoring systems.

ACKNOWLEDGEMENTS

The research was funded by a grant from the Defence, Science and Technology Laboratory (DSTLX1000083208)—an executive government agency that ensures innovative science and technology contribute to the defence and security of the United Kingdom. We thank three anonymous reviewers for their comments and suggestions.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

J.T.M., A.W. and S.J. designed the study. J.T.M. collected the data. J.T.M. and S.J. analysed the data. A.W. and S.J. reviewed the manuscript. J.T.M. wrote the manuscript.

PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1111/ejn.15585>.

DATA AVAILABILITY STATEMENT

All code and data that support the findings of this study are openly available under the Creative Commons Attribution 4.0 International license. Analysis scripts and second level data can be accessed via the Open Science Framework (<https://doi.org/10.17605/OSF.IO/YUJW6>). Raw eye tracking data (edf and ascii files) are openly available at figshare (<https://doi.org/10.6084/m9.figshare.17317886.v1>).

ORCID

Joel T. Martin  <https://orcid.org/0000-0002-4475-3835>
Stephen J. Johnston  <https://orcid.org/0000-0001-9360-8856>

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*(2014), 1–20. <https://doi.org/10.1167/14.4.1>
- Ariga, A., & Lleras, A. (2011). Brief and rare mental “breaks” keep you focused: Deactivation and reactivation of task goals preempt vigilance decrements. *Cognition*, *118*(3), 439–443. <https://doi.org/10.1016/j.cognition.2010.12.007>
- Aston-Jones, G., Chiang, C., & Alexinsky, T. (1991). Discharge of noradrenergic locus coeruleus neurons in behaving rats and monkeys suggests a role in vigilance. *Progress in Brain Research*, *88*(C), 501–520. [https://doi.org/10.1016/S0079-6123\(08\)63830-3](https://doi.org/10.1016/S0079-6123(08)63830-3)
- Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*, 403–450. <https://doi.org/10.1146/annurev.neuro.28.061604.135709>
- Aston-Jones, G., Rajkowski, J., & Cohen, J. D. (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](https://doi.org/10.1016/S0006-3223(99)00140-7)
- Aston-Jones, G., Rajkowski, J., Kubiak, P., & Alexinsky, T. (1994). Locus coeruleus neurons in monkey are selectively activated by attended cues in a vigilance task. *The Journal of Neuroscience*, *14*(July), 4467–4480. <https://doi.org/10.1523/JNEUROSCI.14-07-04467.1994>
- Baddeley, A. D., & Colquhoun, W. P. (1969). Signal probability and vigilance: A reappraisal of the “signal-rate” effect. *British Journal of Psychology*, *60*(2), 169–178. <https://doi.org/10.1111/j.2044-8295.1969.tb01189.x>
- Ballard, J. C. (1996). Computerized assessment of sustained attention: A review of factors affecting vigilance performance. *Journal of Clinical and Experimental Neuropsychology*, *18*(6), 843–863. <https://doi.org/10.1080/01688639608408307>
- Barbur, J. L., Harlow, A. J., & Sahraie, A. (1992). Pupillary responses to stimulus structure, colour and movement. *Ophthalmic & Physiological Optics*, *12*(2), 137–141. <https://doi.org/10.1111/j.1475-1313.1992.tb00276.x>
- Basner, M., & Dinges, D. F. (2011). Maximizing sensitivity of the psychomotor vigilance test (PVT) to sleep loss. *Sleep*, *34*(5), 581–591. <https://doi.org/10.1093/sleep/34.5.581>
- Basner, M., Mollicone, D., & Dinges, D. F. (2011). Validity and sensitivity of a brief psychomotor vigilance test (PVT-B) to total and partial sleep deprivation. *Acta Astronautica*, *69*(11–12), 949–959. <https://doi.org/10.1016/j.actaastro.2011.07.015>
- Beatty, J. (1982). 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>
- Berridge, C. W. (2008). Noradrenergic modulation of arousal. *Brain Research Reviews*, *58*(1), 1–17. <https://doi.org/10.1016/j.brainresrev.2007.10.013>
- Berridge, C. W., Schmeichel, B. E., & España, R. A. (2012). Noradrenergic modulation of wakefulness/arousal. *Sleep Medicine Reviews*, *16*(2), 187–197. <https://doi.org/10.1016/j.smrv.2011.12.003>
- Berridge, C. W., & Waterhouse, B. D. (2003). The locus coeruleus-noradrenergic system: Modulation of behavioral state and state-dependent cognitive processes. *Brain Research Reviews*, *42*(1), 33–84. [https://doi.org/10.1016/S0165-0173\(03\)00143-7](https://doi.org/10.1016/S0165-0173(03)00143-7)
- Bills, A. G. (1943). *The psychology of efficiency: A discussion of the hygiene of mental work*. Harper & Bros.
- Blatter, K., Graw, P., Münch, M., Knoblauch, V., Wirz-Justice, A., & Cajochen, C. (2006). Gender and age differences in psychomotor vigilance performance under differential sleep pressure conditions. *Behavioural Brain Research*, *168*(2), 312–317. <https://doi.org/10.1016/j.bbr.2005.11.018>
- Broadbent, D. E. (1953). Noise, paced performance and vigilance tasks. *British Journal of Psychology*, *44*(4), 295–303. <https://doi.org/10.1111/j.2044-8295.1953.tb01210.x>
- Broadbent, D. E. (1958). *Perception and communication*. Oxford University Press. <https://doi.org/10.1037/10037-000>
- Broadbent, D. E. (1971). *Decision and stress*. Academic Press.
- Broadbent, D. E., & Gregory, M. (1965). Effects of noise and of signal rate upon vigilance analysed by means of decision theory. *Human Factors*, *7*, 155–162. <https://doi.org/10.1177/001872086500700207>
- Buck, L. (1966). Reaction time as a measure of perceptual vigilance. *Psychological Bulletin*, *65*(5), 291–304. <https://doi.org/10.1037/h0023207>
- Caggiano, D. M., & Parasuraman, R. (2004). The role of memory representation in the vigilance decrement. *Psychonomic Bulletin & Review*, *11*(5), 932–937. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1351035&tool=pmcentrez&rendertype=abstract>, <https://doi.org/10.3758/BF03196724>
- Caldwell, J. A., Prazinko, B., & Caldwell, J. L. (2003). Body posture affects electroencephalographic activity and psychomotor vigilance task performance in sleep-deprived subjects. *Clinical Neurophysiology*, *114*(1), 23–31. [https://doi.org/10.1016/S1388-2457\(02\)00283-3](https://doi.org/10.1016/S1388-2457(02)00283-3)
- Colquhoun, W. V. (1961). The effect of ‘unwanted’ signals on performance in a vigilance task. *Ergonomics*, *4*(1), 41–51. <https://doi.org/10.1080/00140136108930506>
- Costa, V. D., & Rudebeck, P. H. (2016). More than meets the eye: The relationship between pupil size and locus coeruleus activity. *Neuron*, *89*(1), 8–10. <https://doi.org/10.1016/j.neuron.2015.12.031>
- de Gee, J. W., Colizoli, O., Kloosterman, N. A., Knapen, T., Nieuwenhuis, S., & Donner, T. H. (2017). Dynamic modulation of decision biases by brainstem arousal systems. *eLife*, *6*(Lc), 1–36. <https://doi.org/10.7554/eLife.23232>
- de Gee, J. W., Knapen, T., & Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences USA*, *111*(5), E618–E625. <https://doi.org/10.1073/pnas.1317557111>
- Dinges, D. F., Pack, F., Williams, K., Gillen, K. A., Powell, J. W., Ott, G. E., Aptowicz, C., & Pack, A. I. (1997). Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4–5 hours per night. *Sleep*, *20*(4), 267–277. <http://www.ncbi.nlm.nih.gov/pubmed/9231952>
- Dorrian, J., Lamond, N., Holmes, A. L., Burgess, H. J., Roach, G. D., Fletcher, A., & Dawson, D. (2003). The ability to self-monitor performance during a week of simulated night shifts. *Sleep*, *26*(7), 871–877. <https://doi.org/10.1093/sleep/26.7.871>

- Dorrian, J., Roach, G. D., Fletcher, A., & Dawson, D. (2007). Simulated train driving: Fatigue, self-awareness and cognitive disengagement. *Applied Ergonomics*, *38*(2), 155–166. <https://doi.org/10.1016/j.apergo.2006.03.006>
- Einhäuser, W., Koch, C., & Carter, O. L. (2010). Pupil dilation betrays the timing of decisions. *Frontiers in Human Neuroscience*, *4*(February), 18. <https://doi.org/10.3389/fnhum.2010.00018>
- Einhäuser, W., Stout, J., Koch, C., & Carter, O. L. (2008). Pupil dilation reflects perceptual selection and predicts subsequent stability in perceptual rivalry. *Proceedings of the National Academy of Sciences USA*, *105*(5), 1704–1709. <https://doi.org/10.1073/pnas.070727105>
- Esterman, M., Grosso, M., Liu, G., Mitko, A., Morris, R., & DeGutis, J. (2016). Anticipation of monetary reward can attenuate the vigilance decrement. *PLoS ONE*, *11*(7), e0159741. <https://doi.org/10.1371/journal.pone.0159741>
- 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*, 70–91. <https://doi.org/10.1111/nyas.13318>
- Franklin, M. S., Broadway, J. M., Mrazek, M. D., Smallwood, J., & Schooler, J. W. (2013). Window to the wandering mind: Pupillometry of spontaneous thought while reading. *The Quarterly Journal of Experimental Psychology*, *66*(12), 2289–2294. <https://doi.org/10.1080/17470218.2013.858170>
- Frankmann, J. P., & Adams, J. A. (1962). Theories of vigilance. *Psychological Bulletin*, *59*(4), 257–272. <https://doi.org/10.1037/h0046142>
- Fried, M., Tsitsiashvili, E., Bonne, Y. S., Sterkin, A., Wagnanski-Jaffe, T., Epstein, T., & Polat, U. (2014). ADHD subjects fail to suppress eye blinks and microsaccades while anticipating visual stimuli but recover with medication. *Vision Research*, *101*, 62–72. <https://doi.org/10.1016/j.visres.2014.05.004>
- 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>
- Goldwater, B. C. (1972). Psychological significance of pupillary movements. *Psychological Bulletin*, *77*(5), 340–355. <http://psycnet.apa.org/journals/bul/77/5/340/>, <https://doi.org/10.1037/h0032456>
- Grandchamp, R., Braboszcz, C., & Delorme, A. (2014). Oculometric variations during mind wandering. *Frontiers in Psychology*, *5*(FEB), 1–10. <https://doi.org/10.3389/fpsyg.2014.00031>
- Graw, P., Kräuchi, K., Knoblauch, V., Wirz-Justice, A., & Cajochen, C. (2004). Circadian and wake-dependent modulation of fastest and slowest reaction times during the psychomotor vigilance task. *Physiology and Behavior*, *80*(5), 695–701. <https://doi.org/10.1016/j.physbeh.2003.12.004>
- Green, D. M., & Swets, J. A. (1974). *Signal detection theory and psychophysics* (Vol. 1). RF Krieger.
- Hockey, R. J. (2013). *The psychology of fatigue: Work, effort and control*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139015394>
- Hopstaken, J. F., van der Linden, D., Bakker, A. B., & Kompier, M. A. J. (2015a). 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>
- Hopstaken, J. F., van der Linden, D., Bakker, A. B., & Kompier, M. A. J. (2015b). 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., & Leung, Y. K. (2016). Shifts in attention during mental fatigue: Evidence from subjective, behavioral, physiological, and eye-tracking data. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(6), 878–889. <https://doi.org/10.1037/xhp0000189>
- Horne, J. A., Brass, C. G., & Pettitt, A. N. (1980). Circadian performance differences between morning and evening “types”. *Ergonomics*, *23*(January), 29–36. <https://doi.org/10.1080/00140138008924715>
- Hupé, J.-M., Lamirel, C., & Lorenceau, J. (2009). Pupil dynamics during bistable motion perception. *Journal of Vision*, *9*(7), 10. <https://doi.org/10.1167/9.7.10>
- 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>
- 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>
- Kahneman, D. (1973). *Attention and effort*. Prentice-Hall. <https://doi.org/10.2307/1421603>
- Kahneman, D., & Treisman, A. M. (1984). Changing views of attention and automaticity. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 29–61). Academic Press.
- Knapen, T., de Gee, J. W., Brascamp, J., Nuiten, S., Hoppenbrouwers, S., & Theeuwes, J. (2016). Cognitive and ocular factors jointly determine pupil responses under equillumiance. *PLoS ONE*, *11*(5), e0155574. <https://doi.org/10.1371/journal.pone.0155574>
- Kristjansson, S. D., Stern, J. A., Brown, T. B., & Rohrbaugh, J. W. (2009). Detecting phasic lapses in alertness using pupillometric measures. *Applied Ergonomics*, *40*(6), 978–986. <https://doi.org/10.1016/j.apergo.2009.04.007>
- Laeng, B., Sirois, S., & Gredeback, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, *7*(1), 18–27. <https://doi.org/10.1177/1745691611427305>
- 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>
- Leys, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. <https://doi.org/10.1016/j.jesp.2013.03.013>
- Loh, S., Lamond, N., Dorrian, J., Roach, G. D., & Dawson, D. (2004). The validity of psychomotor vigilance tasks of less than 10-minute duration. *Behavior Research Methods, Instruments, &*

- Computers, 36(2), 339–346. <https://doi.org/10.3758/BF03195580>
- Mackie, R. R. (1987). Vigilance research--are we ready for counter-measures? *Human Factors*, 29(6), 707–723. <https://doi.org/10.1177/001872088702900610>
- Mackworth, J. F. (1970). *Vigilance and attention*. Penguin.
- Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *The Quarterly Journal of Experimental Psychology*, 1(1), 6–21. <https://doi.org/10.1080/17470214808416738>
- Mackworth, N. H. (1950). Researches on the measurement of human performance. H. M. S. O.
- Manousakis, J., Maccora, J., Ftouni, S., & Anderson, C. (2017). It's in the eyes: Pupillary response provides a physiological marker of alertness and performance impairment. *Journal of Sleep Research*, 26(Suppl. 1), 28. https://doi.org/10.1111/jsr.59_12618
- Maris E., & Oostenveld R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1), 177–190. <https://doi.org/10.1016/j.jneumeth.2007.03.024>
- Martin, J. T., Whittaker, A. H., & Johnston, S. J. (2020). Component processes in free-viewing visual search: Insights from fixation-aligned pupillary response averaging. *Journal of Vision*, 20(7), 5. <https://doi.org/10.1167/jov.20.7.5>
- Massar, S. A. A., Lim, J., Sasmita, K., & Chee, M. W. L. (2016). Rewards boost sustained attention through higher effort: A value-based decision making approach. *Biological Psychology*, 120, 21–27. <https://doi.org/10.1016/j.biopsycho.2016.07.019>
- Mathôt, S., Fabius, J. H., van Heusden, E., & van der Stigchel, S. (2018). Safe and sensible baseline correction of pupil-size data. *Behavior Research Methods*, April, 1–25. <https://doi.org/10.3758/s13428-017-1007-2>
- Mathôt, S., Siebold, A., Donk, M., Vitu, F., & Siebold, A. (2015). Large pupils predict goal-driven eye movements. *Journal of Experimental Psychology: General*, 144, 513–521. <https://doi.org/10.1037/a0039168>
- Matsangas, P., Shattuck, N. L., & Brown, S. (2016). Preliminary validation study of the 3-min wrist-worn psychomotor vigilance test. *Behavior Research Methods*, 1–10, 1792–1801. <https://doi.org/10.3758/s13428-016-0821-2>
- McClelland, L. E., Pilcher, J. J., & Moore, D. D. (2010). Oculomotor measures as predictors of performance during sleep deprivation. *Aviation, Space, and Environmental Medicine*, 81(9), 833–842. <https://doi.org/10.3357/ASEM.2653.2010>
- McIntire, L. K., McKinley, R. A., & Goodyear, C. (2014). Detection of vigilance performance with pupillometry. *Applied Proceedings of the Symposium on Eye Tracking Research and Applications*, 167–174. <https://doi.org/10.1145/2578153.2578177>
- Mittner, M., Boekel, W., Tucker, A. M., Turner, B. M., Heathcote, A., & Forstmann, B. U. (2014). When the brain takes a break: A model-based analysis of mind wandering. *The Journal of Neuroscience*, 34(49), 16286–16295. <https://doi.org/10.1523/JNEUROSCI.2062-14.2014>
- 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*, 4154, 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., van Moort, M. L., & Nieuwenhuis, S. (2016). The pupillary orienting response predicts adaptive behavioral adjustment after errors. *PLoS ONE*, 11(3), 1–11. <https://doi.org/10.1371/journal.pone.0151763>
- 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>
- Parasuraman, R. (1979). Memory load and event rate control sensitivity decrements in sustained attention. *Science*, 205(4409), 924–927. <https://doi.org/10.1126/science.472714>
- Parasuraman, R., & Davies, D. R. (1976). Decision theory analysis of response latencies in vigilance. *Journal of Experimental Psychology: Human Perception and Performance*, 2(4), 578–590. <https://doi.org/10.1037/0096-1523.2.4.578>
- Parasuraman, R., & Davies, D. R. (1982). *The psychology of vigilance*. Academic Press.
- Parasuraman, R., Warm, J. S., & See, J. E. (1998). Brain systems of vigilance. In R. Parasuraman (Ed.), *The attentive brain* (pp. 221–256). The MIT Press. <http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1998-07668-010&site=ehost-live>
- Petersen, S. E., & Posner, M. I. (2012). The attention system of the human brain: 20 years after. *Annual Review of Neuroscience*, 35(1), 73–89. <https://doi.org/10.1146/annurev-neuro-062111-150525>
- Phillips, M. A., Szabadi, E. U., & Bradshaw, C. M. (2000). Comparison of the effects of clonidine and yohimbine on spontaneous pupillary fluctuations in healthy human volunteers. *Psychopharmacology*, 150(1), 85–89. <https://doi.org/10.1007/s002130000398>
- Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, 13, 25–42. <https://doi.org/10.1146/annurev.ne.13.030190.000325>
- Rajkowski, J., Kubiak, P., & Aston-Jones, G. (1993). Correlations between locus coeruleus (LC) neural activity, pupil diameter and behavior in monkey support a role of LC in attention. *Society for Neuroscience - Abstracts*, 19, 47.
- Rajkowski, J., Kubiak, P., & Aston-Jones, G. (1994). Locus coeruleus activity in monkey: Phasic and tonic changes are associated with altered vigilance. *Brain Research Bulletin*, 35(5–6), 607–616. [https://doi.org/10.1016/0361-9230\(94\)90175-9](https://doi.org/10.1016/0361-9230(94)90175-9)
- Ralph, B. C. W., Onderwater, K., Thomson, D. R., & Smilek, D. (2016). Disrupting monotony while increasing demand: Benefits of rest and intervening tasks on vigilance. *Psychological Research*, 81(1948), 432–444. <https://doi.org/10.1007/s00426-016-0752-7>
- Richer, F., & Beatty, J. (1985). Pupillary dilations in movement preparation and execution. *Psychophysiology*, 22(2), 204–207. <https://doi.org/10.1111/j.1469-8986.1985.tb01587.x>
- Ross, H. A., Russell, P. N., & Helton, W. S. (2014). Effects of breaks and goal switches on the vigilance decrement. *Experimental*

- Brain Research*, 232(6), 1729–1737. <https://doi.org/10.1007/s00221-014-3865-5>
- Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology*, 56(6), 1–8. <https://doi.org/10.1111/psyp.13335>
- Smallwood, J., Brown, K. S., Baird, B., Mrazek, M. D., Franklin, M. S., & Schooler, J. W. (2012). Insulation for daydreams: A role for tonic norepinephrine in the facilitation of internally guided thought. *PLoS ONE*, 7(4), e33706. <https://doi.org/10.1371/journal.pone.0033706>
- 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>
- Smallwood, J., Davies, J. B., Heim, D., Finnigan, F., Sudberry, M., O'Connor, R., & Obonsawin, M. (2004). Subjective experience and the attentional lapse: Task engagement and disengagement during sustained attention. *Consciousness and Cognition*, 13(4), 657–690. <https://doi.org/10.1016/j.concog.2004.06.003>
- Smallwood, J., & Schooler, J. W. (2006). The restless mind. *Psychology of Consciousness: Theory, Research and Practice*, 132(6), 946–958. <https://doi.org/10.1037/0033-2909.132.6.946>
- SR Research Ltd. (2010). EyeLink user manual. Version 1.5.2.
- Thomson, D. R., Besner, D., & Smilek, D. (2015). A resource-control account of sustained attention: Evidence from mind wandering and vigilance paradigms. *Perspectives on Psychological Science*, 10(1), 82–96. <https://doi.org/10.1177/1745691614556681>
- Thorne, D. R., Johnson, D. E., Redmond, D. P., Sing, H. C., Belenky, G., & Shapiro, J. M. (2005). The Walter Reed palm-held psychomotor vigilance test. *Behavior Research Methods*, 37(1), 111–118. <https://doi.org/10.3758/BF03206404>
- Unsworth, N., Miller, A. L., & Robison, M. K. (2020). Individual differences in lapses of sustained attention: Oculometric indicators of intrinsic alertness. *Journal of Experimental Psychology: Human Perception and Performance*, 46(6), 569–592. <https://doi.org/10.1037/xhp0000734>
- Unsworth, N., & Robison, M. K. (2016). Pupillary correlates of lapses of sustained attention. *Cognitive, Affective, & Behavioral Neuroscience*, 16(April), 601–615. <https://doi.org/10.3758/s13415-016-0417-4>
- Unsworth, N., & Robison, M. K. (2018). Tracking arousal state and mind wandering with pupillometry. *Cognitive, Affective, & Behavioral Neuroscience*, 18(4), 638–664. <https://doi.org/10.3758/s13415-018-0594-4>
- Unsworth, N., Robison, M. K., & Miller, A. L. (2019). Individual differences in baseline oculometrics: Examining variation in baseline pupil diameter, spontaneous eye blink rate, and fixation stability. *Cognitive, Affective, & Behavioral Neuroscience*, 19(4), 1074–1093. <https://doi.org/10.3758/s13415-019-00709-z>
- Urai, A. E., Braun, A., & Donner, T. H. (2017). Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature Communications*, 8, 14637. <https://doi.org/10.1038/ncomms14637>
- van den Brink, R. L., Murphy, P. R., & Nieuwenhuis, S. (2016). Pupil diameter tracks lapses of attention. *PLoS ONE*, 11(10), 1–16. <https://doi.org/10.1371/journal.pone.0165274>
- van Dongen, H. P. A., & Dinges, D. F. (2005). Sleep, circadian rhythms, and psychomotor vigilance. *Clinics in Sports Medicine*, 24(2), 237–249. <https://doi.org/10.1016/j.csm.2004.12.007>
- van Dongen, H. P. A., Price, N. J., Mullington, J. M., Szuba, M. P., Kapoor, S. C., & Dinges, D. F. (2001). Caffeine eliminates psychomotor vigilance deficits from sleep inertia. *Sleep*, 24(7), 813–819. <https://doi.org/10.1093/sleep/24.7.813>
- van Orden, K. F., Jung, T. P., & Makeig, S. (2000). Combined eye activity measures accurately estimate changes in sustained visual task performance. *Biological Psychology*, 52(3), 221–240. [https://doi.org/10.1016/S0301-0511\(99\)00043-5](https://doi.org/10.1016/S0301-0511(99)00043-5)
- Varazzani, C., San-Galli, A., Gilardeau, S., & Bouret, S. (2015). Noradrenaline and dopamine neurons in the reward/effort trade-off: A direct electrophysiological comparison in behaving monkeys. *The Journal of Neuroscience*, 35(20), 7866–7877. <https://doi.org/10.1523/JNEUROSCI.0454-15.2015>
- Warm, J. S., & Jerison, H. J. (1984). The psychophysics of vigilance. In J. S. Warm (Ed.), *Sustained attention in human performance* (pp. 15–59). John Wiley & Sons.
- Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors*, 50(3), 433–441. <https://doi.org/10.1518/001872008X312152>
- Wiener, E. L. (1987). Application of vigilance research: Rare, medium, or well done? *Human Factors*, 29(6), 725–736. <https://doi.org/10.1177/001872088702900611>
- Wilkinson, R. T., & Houghton, D. (1982). Field test of arousal: A portable reaction timer with data storage. *Human Factors*, 24(4), 487–493. <https://doi.org/10.1177/001872088202400409>
- 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>
- Yoo, K., Ahn, J., & Lee, S.-H. (2021). The confounding effects of eye blinking on pupillometry, and their remedy. *PLoS ONE*, 16(12), e0261463. <https://doi.org/10.1371/journal.pone.0261463>
- Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, 46(4), 681–692. <https://doi.org/10.1016/j.neuron.2005.04.026>

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: 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>