Attention, Perception, & Psychophysics (2019) 81:407–419
https://doi.org/10.3758/s13414-018-1618-4

Pupillometry tracks fluctuations in working memory performance

Matthew K. Robison[1] · **Nash Unsworth**[2]

Published online: 8 November 2018

Abstract

In 3 experiments, we examined fluctuations in working memory (WM) performance and associated changes in pretrial and task-evoked pupil diameter. Additionally, we examined whether particularly poor trials were accompanied by self-reports of off-task attentional states. The results demonstrated that task-evoked pupillary responses can be used to measure moment-to-moment fluctuations in the success of WM maintenance during delay intervals. Further, when individuals reported being in an off-task attentional state, their WM performance suffered. Additionally, when probed directly after a particularly poor trial, participants reported being in an off-task attentional state more often than at random intervals throughout the task. So behavioral, subjective, and physiological data converged when people experienced WM failures. Although pretrial pupil diameter did not consistently differentiate between successful and unsuccessful trials, variability in pretrial pupil diameter accounted for a significant portion of variance in WM task performance. This effect persisted after controlling for mean task-evoked pupillary response and variability in task-evoked pupillary responses. Thus, one of the major reasons people varied in the consistency with which they utilized their WM system was variability in arousal. Such variability in arousal is potentially due to variation in the functioning of the locus coeruleus-norepinephrine (LC-NE) neuromodulatory system, and thus may underlie individual differences in WM capacity and attention control.

Keywords Working memory · Attention · Pupillometry · Mind-wandering

Working memory (WM) is a capacity-limited system that temporarily maintains and manipulates information. The precise nature of these limitations is still a subject of debate, but individuals can only retain a few pieces of information at a time, typically around three or four items (Cowan, 2001; Luck & Vogel, 1997). Furthermore, individual differences in WM capacity substantially correlate with other cognitive abilities, such as attention control, long-term memory, reading comprehension, and fluid intelligence (Daneman & Carpenter, 1980; Engle et al., 1999; Rosen & Engle, 1997; Unsworth et al., 2014; Unsworth & Spillers, 2010). Recent research has shown that one important source of variance in WM capacity is not necessarily the amount of information one can maintain at one given time, but rather the consistency with which WM capacity is effectively utilized on a moment-to-moment basis (Adam et al., 2015; Unsworth & Robison, 2016b). In the present study, we focus on fluctuations in WM and how they might be measured with behavior, subjective reporting of attentional state, and pupillometry.

A recent study demonstrated that the contralateral delay activity (CDA), a sustained negativity that occurs over occipital and parietal scalp electrode sites in EEG, tracks fluctuations in WM performance (Adam et al., 2018). The magnitude of CDA has previously been demonstrated to correlate with individual differences in WM capacity and other cognitive abilities (Unsworth et al., 2015; Vogel et al., 2005). Adam et al. (2015) utilized a discrete whole-report WM task, which requires participants to report all items in a memory array, rather than a single item, like in typical change-detection tasks (e.g., Luck & Vogel, 1997). Rather than averaging over a number of trials to get an estimate of the individual's WM capacity, whole-report tasks track how much information the individual effectively encodes and maintains on a trial-by-trial basis. For example,

All data and analysis scripts, including the R Markdown script used to generate this manuscript, are available on the Open Science Framework at the unique URL: https://osf.io/vuw9h/.

✉ Matthew K. Robison
matthewkrobison@asu.edu

[1] Department of Psychology, Arizona State University, 950 S. McAllister Ave., Tempe, AZ 85287, USA

[2] Department of Psychology, University of Oregon, Eugene, OR, USA

when given six items to remember, participants sometimes remember only a very small amount of information (e.g., one item) or no information at all (Adam et al. 2015, 2018; Adam and Vogel, 2016, 2017). Thus, people do not always fully utilize their WM capacity. Adam et al. (2015) argued these instances are due to lapses in attention, which cause information to either not be encoded properly into the WM system or lost during a maintenance interval. When measuring EEG during such a task, Adam et al. (2018) found significantly higher CDA amplitude when performance was relatively good (four or more items correctly reported) compared to when performance was relatively poor (two or fewer items correctly reported). Thus, the CDA can be used to track the contents of WM on a moment-to-moment basis.

In the present study, we sought to extend these findings with a different physiological measure of WM storage—pupillometry—and to corroborate the idea that these instances do indeed constitute lapses of attention. Previously, pupillometry has been used to study memory processes in a number of ways. For example, Kahneman and Beatty (1966) showed that the pupil dilates as people store information in WM, and the magnitude of the dilation scales with how much information people are required to remember. These findings have been replicated rather recently with complex span WM tasks (Heitz et al., 2008). Other investigations have shown that pupil diameter dilates in response to tracking multiple objects (Alnæs et al., 2014), that pupil dilation at encoding distinguishes between high- and low-strength memories (Papesh et al., 2012), and that pupil dilation at encoding is sensitive to the value of to-be-remembered information in a value-directed remembering paradigm (Ariel & Castel, 2014). In our own work, we have demonstrated that the pupil can be used to index the storage of information in WM during a delay in a variety of WM task situations (Unsworth & Robison, 2018b), and that pupillometry can be used to measure individual differences in the allocation of attention to items in WM (Unsworth & Robison, 2015).

In a separate yet related line of research, we have been investigating the pupillary correlates of attentional lapses within sustained attention tasks (Unsworth & Robison, 2016a, 2018a; Unsworth et al. 2018). In these studies, we have observed differences in both pretrial pupillary dynamics and task-evoked pupil dilations when participants are in relatively disengaged attentional states. Specifically, when people report being in an off-task attentional state (e.g., mind-wandering or "zoning out"), task-evoked pupil dilations were shallower and performance on the task was worse (i.e., reaction times were longer). Further, long reaction times were preceded by reductions in pupil dilation in preparation for the upcoming stimulus (Unsworth et al., 2018). Theoretically, pretrial pupil diameter provides an index of baseline activity in the locus coeruleus-norepinephrine (LC-NE) neuromodulatory system, which is partially responsible for regulating the allocation of attentional resources, maintaining alertness, and attending to task-relevant sources of information in the environment (Aston-Jones & Cohen, 2005; Gilzenrat et al., 2010; Murphy et al., 2011; Sara, 2009). Indeed, we have argued that variation in the consistent functioning of the LC-NE system underlies individual differences in WM capacity and attention-control abilities (Unsworth & Robison, 2017a), and we have demonstrated that variation in pretrial and task-evoked pupil diameter during attention control tasks correlates with individual differences in WM capacity, attention-control abilities, and the tendency to mind-wander during such tasks (Unsworth & Robison, 2017b). Therefore, both pretrial pupil diameter and task-evoked pupil diameter may potentially measure fluctuations in performance in a WM task. There is some preliminary evidence to suggest that pretrial pupil diameter might indicate an upcoming lapse in WM performance. Pretrial pupil diameter was smaller when participants incorrectly responded to small set sizes (1 and 2) in a change-detection task (Unsworth & Robison, 2015) compared to when they accurately responded. However, it is worth noting that participants only received 20 trials each of set size 1 and 2, and performance on these trials was very high (> 90% accuracy) on average. Therefore, these measurements are based on a small number of trials and a smaller number of participants, as some participants never incorrectly responded on such small set sizes.

The present study sought to answer several questions. First, will pupil dilation track fluctuations in WM performance on a trial-by-trial basis? That is, will pupil dilation during delay intervals be greater when participants accurately report relatively more information in a whole-report task, even when the amount of to-be-remembered information is held constant across trials? Further, will differences in pretrial pupil diameter precede relatively poor trials? Second, will "lapses" in behavioral performance (i.e., trials on which people report 0 or 1 correct items) be accompanied by subjective reports that people are in off-task attentional states? If so, this would provide credence to the idea that WM performance below capacity limits represent lapses of attention. Together, these experiments will expand our understanding of WM capacity limitations, the physiological correlates of WM failures, and perhaps provide avenues for mitigating and/or preventing such failures.

Experiment 1

Experiment 1 examined whether pupil diameter could be used to track trial-to-trial fluctuations in WM performance.

Participants completed a visual WM task while their pupil diameters were continuously recorded.

Method

Participants and procedure

Our target sample size was 30 participants, which was based on our prior investigations of WM using pupillometry (Unsworth & Robison, 2018b). A sample of 34 participants from the human subjects pool at the University of Oregon completed the study in exchange for partial course credit. We stopped collecting data on the day we reached our target sample size. After completing informed consent and demographic forms, the participants completed a 30-min sustained attention task that was irrelevant to the current study. Participants then completed the discrete whole-report task during the second half of the 1-h session. We debriefed participants at the end of the experimental session. We treated all participants according to the ethical standards of the American Psychological Association.

Task

The participants' task was to remember the colors of squares over brief delays and to report the colors of these squares on a testing screen (Adam et al., 2015, 2018; Adam & Vogel, 2016, 2017). Each trial began with a 1-s fixation screen on which a black fixation crossed appeared on a grey background, followed by a 100-ms blank screen. Then, a pattern of six colored squares appeared and remained on screen for 250 ms. The squares (60 × 60 pixels; 3° visual angle) appeared within a 540 × 402-pixel region centered on the screen. The locations were random with the restriction that no items appeared within a 100-pixel vector distance of each other (measured from each item's top-left starting point). Colors were randomly sampled from a set of nine discrete colors (white, black, red, blue, lime green, magenta, green, cyan, and yellow). Colors did not repeat within a trial (i.e., all six items were different colors). After a 3900-ms blank delay screen, the color response grids appeared in the locations where the six items had appeared previously. The participants' task was to report the color of the square in each location by clicking the appropriate color in the grid. After the participant responded to all six items, the next trial immediately started. Figure 1 presents a graphical depiction of the task. Participants first read through a series of instruction screens followed by five practice trials. If participants were confused during the practice trials, they were encouraged to seek clarification from the experimenter (the first author). They then completed 80 experimental trials.

Pupil measurement

Participants sat with their heads in a chin rest at a distance of 60 cm from the monitor. Pupil data were continuously measured via a Tobii T300 eye-tracker at a sampling rate of 120 Hz. The eye-tracker measured pupil diameter from both eyes in millimeters, so no conversion was necessary. We used the left eye's pupil diameter for all analyses (right and left eye pupil diameter correlated at r = .88). The eye-tracker was calibrated to record the position of each eye in x-y coordinates, but because we did not have any predictions regarding eye-movements and specified no areas of interest a priori, we do not examine or report eye position or saccades in the present study. Participants were instructed to fixate on the cross during the 1-s fixation screen. Participants were allowed to move their eyes after that point during the trial. All trials were included, but missing data due to blinks and off-screen fixations were excluded from the analyses.

Thought probes

After the practice trials, the instructions told participants, "We are also interested in finding out how often your mind wanders and you are distracted during tasks like this. In order to examine this, the computer will periodically ask you what you were just thinking about. It is normal for your mind to wander from time to time on a task like this." The thought probes presented participants with a screen that said, "Please characterize your current conscious experience." The screen listed six options: (1) I am totally focused on the current task, (2) I am thinking about my performance on the task, (3) I am distracted by sights/sounds in my environment, (4) I am intentionally thinking about things unrelated to the task, (5) I am unintentionally thinking about things unrelated to the task, (6) My mind is blank. The participant pressed the key on the keyboard that best described their current thoughts. The next trial immediately began after the participant responded to the thought probe. Due to a programming error, the responses to these thought probes were not recorded in Experiment 1. We corrected this error for Experiments 2 and 3.

Results

The analyses were performed using R software (R Core Team, 2017) and the manuscript was developed using the *papaja* package (Aust & Barth, 2018) and an R markdown script. The interested reader can download the R markdown script and all data files from the Open Science Framework (https://osf.io/vuw9h/). Eye-tracking and behavioral data were preprocessed offline to aggregate to the level of analysis.

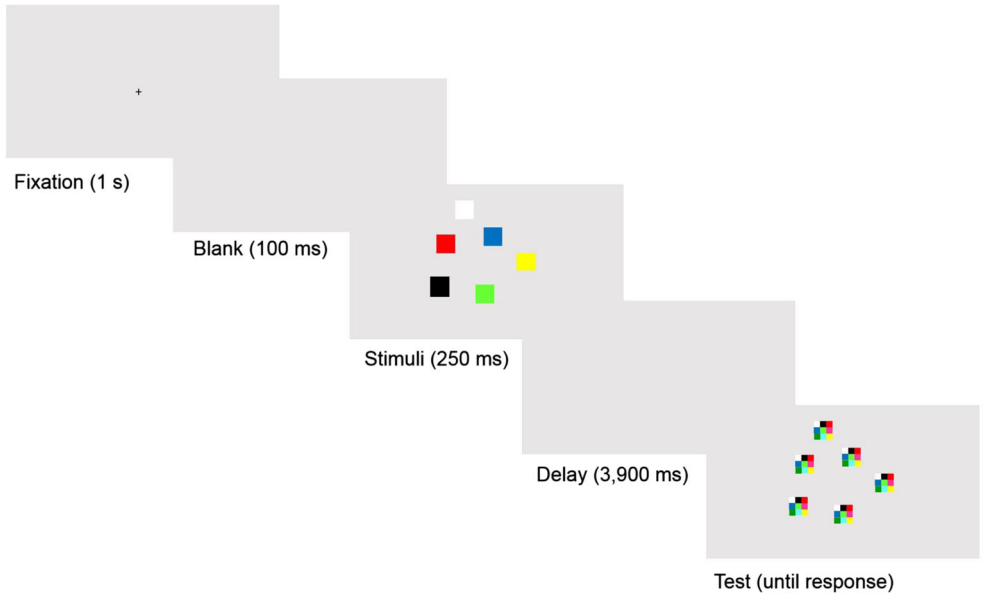

Fig. 1 Discrete whole report task used in all experiments

Task performance

Our first analysis focused on behavioral performance in the discrete whole-report task. Participants correctly reported an average of 2.96 items (SD = 0.46), which is on par with prior research using this task (Adam et al., 2015, 2018; Adam & Vogel, 2016, 2017), and consistent with average estimates of WM capacity (Cowan, 2001; Luck & Vogel, 1997). A distribution of correct items reported is shown in Fig. 2. The modal number correct was three items, but participants also frequently reported two or fewer items correctly, which is presumably below capacity for most people. So participants frequently encountered instances in which they did not maximize their WM capacity.

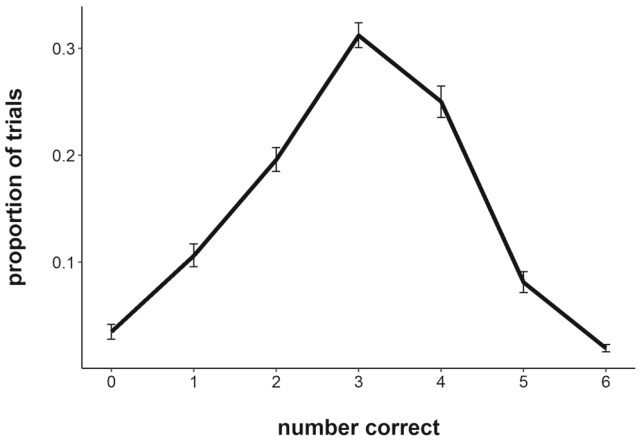

Fig. 2 Distribution of correct responses in Experiment 1. *Error bars* represent ± one standard error of the mean

Pupillometry

To measure pretrial arousal levels, we computed a pretrial pupil diameter for each trial for each individual by averaging over the 1-s fixation screen. To compute attentional allocation during the delay, we averaged pupil measurements into thirty-nine 100-ms bins for each individual and each trial. These measurements were then subtracted from the average of pupil diameter during the 100-ms pre-stimulus blank screen. Thus, the task-evoked response is measured as a change (i.e., dilation) from this baseline. In our first analysis, we ensured that the task-evoked pupil response appeared similar to prior examinations of visual WM using pupillometry (Unsworth & Robison, 2015, 2018b). A repeated-measures analysis of variance (ANOVA) with bin as a within-subjects factor indicated a significant main effect of bin ($F(38, 1216) = 29.38$, $p < 0.01$, partial $\eta^2 = 0.48$). As can be seen in Fig. 3, the pupil dilated during first ˜1500 ms of the delay and sustained this dilation during the majority of the delay.

To examine how pupillometric measures accompanied fluctuations in performance, we categorized "good" trials as any trial in which participants correctly reported four or more items, and "poor" trials as any trial in which participants correctly reported two or fewer items. This separation of good and poor trials was specified based on the distinction made by Adam et al. (2018, 2015), who performed this same analysis using a nearly-identical task but examined how EEG measures (N1, N2PC, CDA, and theta power) differed based on relatively good performance vs. relatively poor performance. Indeed, one of the

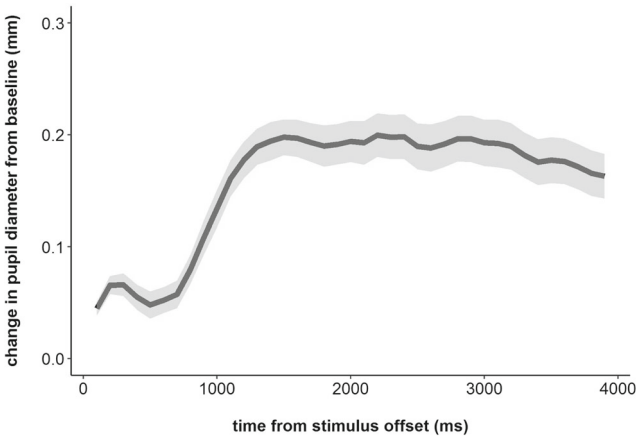

Fig. 3 Average task-evoked pupillary response in Experiment 1. *Shaded error bar* represents ± one standard error of the mean

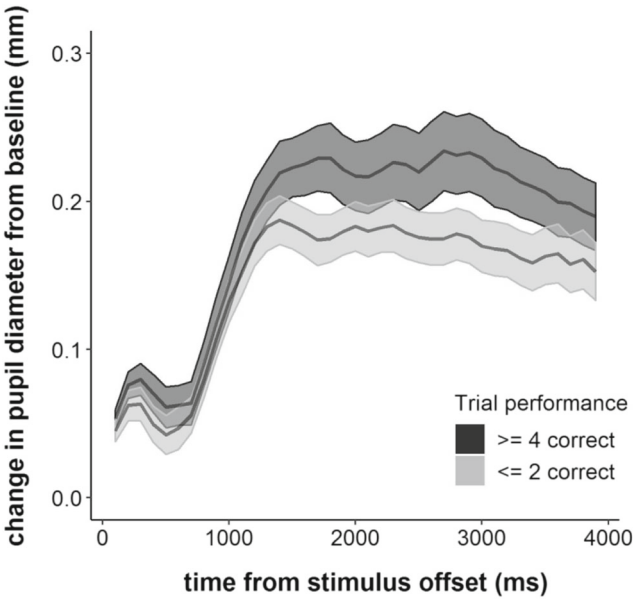

Fig. 4 Task-evoked pupillary response for good and poor trials in Experiment 1. *Shaded error bars* represent ± one standard error of the mean

primary goals of this study was to use this same procedure with pupillometry. On average, participants had 28.03 (*SD* = 11.16) good trials and 27.30 (*SD* = 11.35) poor trials contributing to the analyses. All participants had at least seven good trials and seven poor trials to analyze. Comparing pretrial pupil diameter between good and poor trials revealed a small but significant difference. However, this effect was in the opposite of the hypothesized direction. Average pretrial pupil diameter was significantly larger preceding poor trials (M = 3.33 mm, SD = 0.48) compared to good trials (M = 3.30, SD = 0.47, $t(32)$ = -3.37, $p < 0.01$, Cohen's d = 0.59). To compare task-evoked pupil diameter for good and poor trials, we submitted pupil diameter to a 2 (trial-type: good, poor) x 39 (bin: 1 - 39) repeated-measures ANOVA. The analysis indicated a significant main effect of trial-type ($F(1, 32)$ = 12.02, $p < 0.01$, partial η^2 = 0.27), such that average pupil dilation was larger for good trials, and a bin x trial-type interaction ($F(38, 1216)$ = 3.30, $p < 0.01$, partial η^2 = 0.09), such that the two trial types showed different responses across the delay. As can be seen in Fig. 4, the pupil dilated to a larger extent on trials where participants correctly reported a relatively high number of items. On trials when participants reported fewer items, the pupil dilated to a lesser extent. This is evidence that the pupil tracks the success of WM on a trial-by-trial basis.

Discussion

In Experiment 1, we expected pretrial pupil diameter to be smaller preceding poor trials, which would indicate a low-arousal, inattentive mental state that should lead to failures of WM. However, we did not observe this effect. In fact, we observed a small but significant effect in the opposite direction. Pretrial pupil diameter was larger preceding poor trials compared to good trials. Comparing task-evoked pupillary responses for good and poor trials supported our hypothesis. Although the pupil showed a sustained dilation above baseline during the delay for both good and poor trials, the pupil dilated to a significantly larger extent during good trials. Presumably, the task-evoked pupillary response reflects an effortful allocation of attention to retaining information in WM. On trials where it appeared participants exerted more attention, they accurately reported more items. These results nicely replicate recent findings by Adam et al. (2018), where the CDA tracked WM performance on a trial-by-trial basis. In that study, Adam et al. (2018) argued that that WM failures were caused in part by storage failures. The pupillary differences between good and poor trials in the present study provide further evidence that such failures are potentially due to a lack of effortful attention during the delay period.

Experiment 2

Experiment 2 corrected the programming error made in Experiment 1 wherein thought probe responses were not recorded. Otherwise, the task and procedure were nearly identical to Experiment 1. Thought probe responses allowed us to measure (1) how often people report being in various attention states on average during this task, (2) if self-reports of off-task attention states are accompanied by worse task performance, and (3) if self-reports of off-task attention

states are accompanied by differences in pretrial and/or task-evoked pupil diameter.

Method

Participants and procedure

A sample of 36 participants from the human subjects pool at the University of Oregon completed the study in exchange for partial course credit. Three participants were excluded from the final analysis for abnormally poor performance on the task[1], leaving a final sample of 33 participants. All participants first completed informed consent and demographics forms. They then completed the discrete whole report task during the first half of a 1-h session. During the second half of the session, participants completed a sustained attention task that was irrelevant to the present investigation. We debriefed participants at the end of the experimental session. We treated all participants according to the ethical standards of the American Psychological Association.

Task

The discrete whole report task was identical to that used in Experiment 1. The task included eight thought probes to measure participants' subjective attentional state at various points throughout the task. These thought probes were also included in Experiment 1, but due to a programming error, responses to the probes were not collected. The only other procedural difference was that participants completed the task on a Tobii T120 eye-tracker, rather than a Tobii T300 eye-tracker. Additionally, participants completed the discrete whole report task during the first half of the experimental session, rather than the second half of the session, as in Experiment 1. As noted in the Results section, this did not significantly affect task performance, so we do not believe it was an important confound between experiments. There were slight differences in the magnitude of pretrial and task-evoked pupil diameters. This may have been due to slight differences in background luminance across experimental settings.

Results

Participants correctly reported an average of 2.99 items correctly (SD = 0.60), which was about equal to performance in Experiment 1 ($t(65) = -0.25$, $p = 0.80$, Cohen's $d = -0.06$). The modal number correct was three items, but participants also frequently reported two or fewer items correctly, which is presumably below capacity for most people. So participants frequently encountered instances in which they did not maximize their WM capacity, similar to Experiment 1. Figure 5 shows these data.

Pupillometry

Just as in Experiment 1, we categorized "good" trials as any time a participant reported four or more items correctly, and "poor" trials as any time a participant reported two or fewer items correctly. This analysis included an average of 28.03 (SD = 13.26) good trials and 27.06 (SD = 13.93) poor trials. All participants had at least five good trials and eight poor trials contributing to these analyses. In the first analysis, we examined pretrial pupil diameter by trial performance. As a reminder, in Experiment 1, we observed a significant difference in pretrial pupil diameter between good and poor trials, but in the opposite of the expected direction. This same analysis for Experiment 2 revealed a non-significant difference in pretrial pupil diameter between good and poor trials (M good trials = 3.01 mm, SD = 0.43; M poor trials = 3.00, SD = 0.42; paired-samples $t(32) = 1.87$, $p = 0.07$, Cohen's $d = 0.33$). So although the effect was in the hypothesized direction, the difference was not quite significant.

Next, we submitted binned task-evoked pupil data to a 2 (trial-type: good, poor) x 39 (bin: 1 - 39) repeated-measures ANOVA. The analysis indicated a significant main effect of trial-type ($F(1, 31) = 7.47$, $p = 0.01$, partial $\eta^2 = 0.02$), such that overall task-evoked pupil diameter was greater for good trials compared to poor trials, a main effect of bin ($F(38, 1214) = 25.08$, $p < 0.01$, partial $\eta^2 = 0.19$), and a trial-type x bin interaction ($F(38, 1214) = 4.67$, p

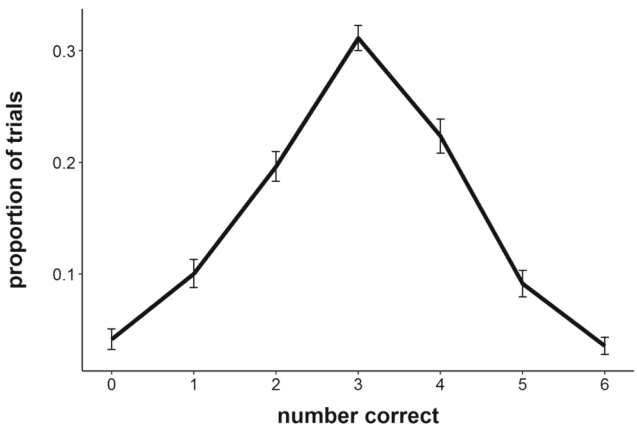

Fig. 5 Distribution of correct responses in Experiment 2. *Error bars* represent ±1 standard error of the mean

[1]These participants correctly reported an average of 1.04, 1.15, and 1.36 items. So it was not clear that they understood the task instructions.

< 0.01, partial $\eta^2 = 0.003$). All of these effects replicated the findings in Experiment 1. Again, pupil dilation during the delay period tracked how much information participants held in memory on a trial-by-trial basis. The pupil dilated to a greater extent when participants reported a relatively large number of items correctly than when they held relatively few items in memory. Figure 6 shows this pattern.

Attention states

Our next set of analyses focused on responses to thought probes. Table 1 shows average proportions for each response. Comparing performance on trials immediately preceding an on-task report to trials immediately preceding any off-task report yielded a significant difference (*M* on-task = 3.41, *SD* = 1.18; *M* off-task = 2.86, *SD* = 1.05; paired-samples $t(17) = 2.82$, $p = 0.01$, Cohen's $d = 0.66$). Thus, when participants were in some sort of off-task state, whether it be mind-wandering, mind-blanking, or experiencing external distraction, they were able to hold fewer items in memory. This replicates prior research (Adam & Vogel, 2017; Unsworth & Robison, 2016b).

The next set of analyses examined pupillary measures with respect to thought probe responses. The first analysis compared pretrial pupil diameter preceding on-task reports vs. off-task reports. This analysis indicated no significant difference between pretrial pupil diameter as a function of attention state (*M* on-task = 3.02, *SD* = 0.46; *M* off-task = 2.97, *SD* = 0.45; paired-samples $t(17) = -0.02$, $p = 0.99$, Cohen's $d = 0.00$). Next, we compared average task-evoked pupil diameter as a function of attention state. Although the effect was in the hypothesized direction, average task-evoked pupil diameter did not differ between on- and off-task attention states (*M* on-task = 0.09, *SD* = 0.13; *M* off-task = 0.07, *SD* = 0.13; paired-samples $t(17) = 0.85$, $p = 0.41$, Cohen's $d = 0.20$). For both of these analyses, we should note that sampling is quite limited. There were only eight thought probes in the task, and only 18 participants reported being in an on-task *and* an off-task state at least once. Further, some of these measurements are based on a single trial for each participant. Thus, the reliability of the measurements may below, and thus any finding (or lack thereof) should be interpreted with a degree of caution.

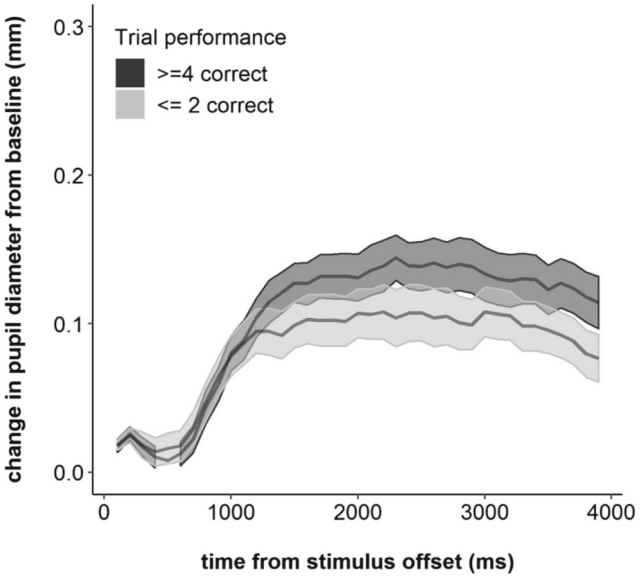

Fig. 6 Task-evoked response for good and poor trials in Experiment 2. *Shaded error bars* represent ±1 standard error of the mean

Table 1 Thought probe response proportions in Experiment 2

	Mean	SD
On task	0.28	0.29
Task-related interference	0.35	0.26
External distraction	0.04	0.14
Intentional mind-wandering	0.04	0.10
Unintentional mind-wandering	0.22	0.19
Mind-blanking	0.07	0.17

SD = standard deviation

Discussion

Experiment 2 had three goals. First, we wanted to replicate the finding that task-evoked pupil responses would track WM performance on a trial-by-trial basis. Second, we wanted to measure how often people fell into off-task attentional states (e.g., mind-wandering) during this task. Third, we wanted to see whether off-task attentional states coincided with worse WM performance. We replicated Experiment 1 such that task-evoked pupillary dilations during the delay period were greater when participants correctly reported a relatively higher number of items. However, we did not find a significant difference in pretrial pupil diameter between good and poor trials. When probed about their thoughts, participants reported being in an off-task attentional state about 37% of the time. Further, on trials immediately preceding such reports, WM performance was significantly worse. There was no significant difference in pretrial pupil diameter preceding poor trials, so this effect did not replicate from Experiment 1. Further, task-evoked pupil diameter preceding reports of off-task attentional states did not differ from that preceding on-task reports. However, because these measurements were based on very few instances per participant, these results should be interpreted with caution.

Experiment 3

The goal of Experiment 3 was to leverage the findings of Experiment 2 to detect attentional lapses. If "lapses" are indeed indicative of subjective states of withdrawn attention to the task, then participants should report being in an off-task subjective state during such trials. To test whether this is the case, we designed the discrete whole-report task to include two types of thought probes. The first type, which we will call "standard probes", appeared after eight random trials. The second type, which we will call "catch" probes, appeared after any trial in which a participant correctly reported 0 or 1 items - so called "lapse" trials (Adam et al., 2015). Note lapse trials are defined slightly differently than "poor" trials were in Experiments 1 and 2. Whereas we used two or fewer correct to denote poor performance based on prior work (Adam et al. 2015, 2018), we used the more stringent standard of 0 or 1 correct to denote lapses. This was done for a couple of reasons. First, we wanted to catch rather extreme instances of task inattention. Second, we wanted to keep the relative frequencies of standard and catch probes about equal. If participants report being in an off-task attentional state (e.g., mind-wandering, mind-blanking, external distraction) more often following the catch probes compared to the randomly delivered standard probes, this would provide evidence that WM failures occur during periods of inattention, and we can indeed call these trials "lapses". If the catch probes do not reveal off-task attentional states, beyond what would be expected by randomly asking participants to report their thoughts, then WM failures may not necessarily be caused by attentional lapses.

Method

Participants and procedure

A sample of 40 participants from the human subjects pool at the University of Oregon participated in the study in exchange for partial course credit. The procedure was nearly identical to Experiment 1. Participants completed a 30-min sustained attention task during the first half of the session and the discrete whole-report task during the second half of the session. No participants from Experiment 1 or Experiment 2 participated in Experiment 3. Two participants were excluded from the final analysis for particularly poor task performance[2], leaving a final sample of 38 participants.

[2]These participants correctly reported an average of 1.30 and 1.21 items correctly. So it was not clear that they understood the task instructions.

Task

The task was nearly identical to that used in Experiment 2 with one exception. In addition to eight "standard" thought probes that appeared after a random sample of trials, we included "catch" probes that appeared after any trial in which a participant correctly reported 0 or 1 items.

Results

On average, participants correctly reported 2.99 items (SD = 0.64) items. Performance was roughly equal to Experiments 1 and 2. The distribution of performance was also quite similar to Experiments 1 and 2 (see Fig. 7).

Attention states

The next set of analyses focused on responses to the "standard" thought probes, which occurred after eight random trials, and the "catch" thought probes, which occurred any time a participant correctly reported 0 or 1 items. On average, participants encountered 10.18 of these instances (SD = 8.84), and only one participant never encountered one. This participant was excluded from analyses comparing standard and catch probes. Proportions of responses to the thought probes are listed in Table 2. We summed instances of off-task attentional states (external distraction, mind-wandering, and mind-blanking) to compute the proportion of time each participant reported being in such a state. A paired-samples t-test revealed that participants were significantly more likely to report being in an off-task attentional state when they experienced a behavioral "lapse" (M = 0.47, SD = 0.34) than in randomly sampled instances throughout the task (M = 0.32, SD 0.32, $t(36)$ = 3.69, $p < 0.01$, Cohen's d = 0.61). This

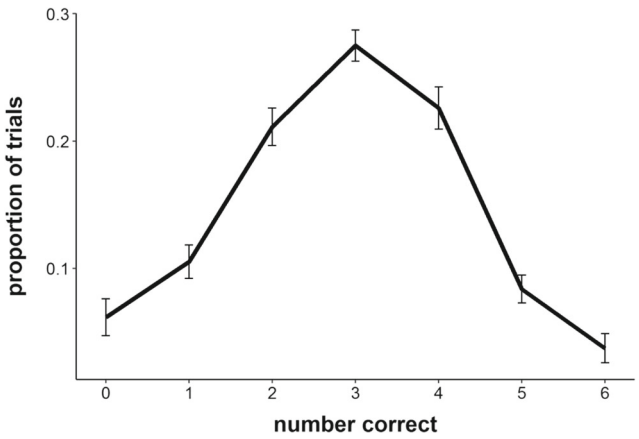

Fig. 7 Distribution of correct responses in Experiment 3. *Error bars* represent ±1 standard error of the mean

Table 2 Thought probe response proportions in Experiment 3

	M catch	SD	M standard	SD
On task	0.27	0.33	0.36	0.34
Task-related interference	0.26	0.23	0.32	0.27
External distraction	0.03	0.09	0.01	0.03
Intentional mind-wandering	0.05	0.14	0.02	0.06
Unintentional mind-wandering	0.20	0.22	0.16	0.18
Mind-blanking	0.18	0.26	0.13	0.27

M = mean, SD = standard deviation

provides evidence that these lapses do often indicate off-task attentional states.

Pupillometry

To examine whether the pupillometry findings from Experiments 1 and 2 replicated, we again analyzed pretrial and task-evoked pupil diameter as a function of task-performance. On average, participants had 28.95 good trials (SD = 15.09) and 28.50 poor trials (SD = 14.39) contributing to the analyses. Comparing pretrial pupil diameter between good trials (M = 2.96, SD = 0.31) and poor trials (M = 2.97, SD = 0.32) revealed no significant difference ($t(37)$ = -1.43, p = 0.16, Cohen's d = -0.23). So Experiment 3 did not replicate the finding from Experiment 1 of different pretrial pupil diameter preceding trials with relatively good WM performance and relatively poor performance.

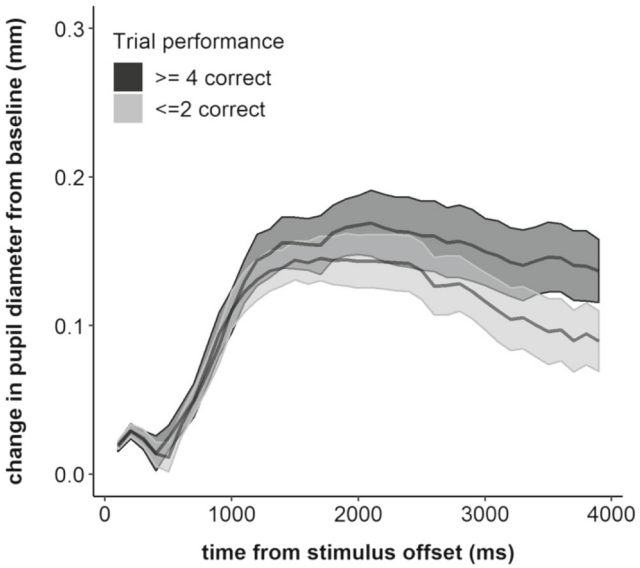

Fig. 8 Task-evoked responses for good and poor trials in Experiment 3. *Error bars* represent ±1 standard error of the mean

Next, we submitted task-evoked pupil diameter to the same analysis as in Experiments 1 and 2. The 2 (trial-type: good, poor) x 39 (bin: 1 - 39) ANOVA indicated a significant main effect of trial-type ($F(1, 37)$ = 4.02, p = 0.05, partial η^2 = 0.02), such that task-evoked responses were larger for good trials compared to poor trials, a main effect of bin ($F(38, 1458)$ = (25.72), $p < 0.01$, partial η^2 = 0.05), and a bin x trial-type interaction ($F(38, 1458)$ = 2.55, $p < 0.01$, partial η^2 = 0.10). This pattern is depicted in Fig. 8. Overall, the task-evoked responses replicated Experiments 1 and 2 in that pupil diameter during the delay interval tracked relative WM performance.

Finally, we examined examine whether differences in task-evoked responses accompanied off-task reports (collapsed across standard and catch thought probes). To do so, we averaged task-evoked responses for on- and off-task reports and compared them. This comparison yielded a non-significant difference in average task-evoked response for on-task reports (M = 0.13, SD = 0.13) and off-task reports (M = 0.11, SD = 0.08, $t(23)$ = -0.37, p = 0.72, Cohen's d = -0.07).

Discussion

Building on Experiment 2, Experiment 3 attempted to "catch" participants in an off-task attentional state using their performance on the immediately preceding trial. Any time a participant encountered a lapse in WM performance (i.e., reported 0 or 1 items correctly) we delivered a thought probe. On average, participants encountered about ten of these instances during the task. Responses to these "catch" thought probes were compared to probes inserted after eight randomly sampled trials. Participants reported being in an off-task attentional state significantly more often following a lapse compared to other random points throughout the task. These data provide evidence that behavioral lapses from such a task are often indicative of off-task attentional states. However, this proportion was still only 47%. Participants reported being on-task 27% of

the time following a lapse and experiencing task-related interference (e.g., thinking about their performance on the task) 26% of the time. So not all behavioral lapses coincided with reports of being subjectively off-task.

WM failures can happen for a number of reasons. Someone can provide too little attention during stimulus presentation, thus failing to encode the items properly, the items can be lost during the maintenance interval, or the participant can have trouble retrieving the items' feature-space bindings at recall. Thus, it is worth noting the time-course of the task and the thought probes. The thought probe sometimes appeared 10 s or more after the appearance of the stimuli. Thus, WM failures due to poor attention at encoding may not be captured by a thought probe that happens later in time. Perhaps by the time the probe appears, the participant has re-engaged with the task. The lack of temporal specificity for the "catch" probe may be why only about half of lapses are followed by an off-task report.

Combined analyses – individual differences

Although we did not have enough power to analyze individual differences in each individual experiment, the combined sample size (N = 107) across the three experiments allowed us to examine individual differences in mean pretrial pupil diameter, variation in pretrial pupil diameter, mean task-evoked pupillary response, variability in task-evoked pupillary responses, and task performance. First, to put participants on the same scale across experiments and across eye-trackers, we standardized mean pretrial pupil diameter and mean task-evoked pupil diameter within each experiment. This standardization procedure involved computing each participant's mean pre-trial pupil diameter, mean task-evoked pupil diameter, and their coefficient of variation in pre-trial and task-evoked pupil diameter by computing their standard deviation across trials and dividing that by their mean. Then, we estimated z-scores for each participant within experiments (i.e., participants in Experiment 1 were treated as one population, Experiment 2 as a separate population, etc.). The number of lapses is simply the sum of trials in which participants reported 0 or 1 items correctly, and mean number correct is each participant's average number of correctly reported items. Table 3 shows correlations among the measures. First, and not surprisingly, the number of lapses strongly and negatively correlated with mean number correct. That is, participants who encountered more instances of extreme inattention to the task performed worse overall (note these measures are partially dependent). Second, while mean pretrial pupil did not significantly correlate with task performance or number of lapses, variability in pretrial pupil diameter did. Finally, both mean task-evoked pupillary response and variability in task-evoked pupillary response also correlated with average performance and number of lapses. In general, people who experienced more fluctuations in arousal performed worse on the task, and people who allocated more attention to the items during the delay periods (and more consistently) performed better. All of these correlations replicate prior work measuring pre-trial pupil diameter, task-evoked pupillary responses, intra-individual variability in these measures, and WM capacity (Unsworth et al. 2015, 2017b, 2018b).

To see if these measures accounted for shared or independent sources of variance in task performance, we regressed mean pretrial pupil diameter, pretrial pupil variability, mean task-evoked response, and task-evoked response variability on mean number correct and number of lapses in separate multiple regressions. Tables 4 and 5 show the results of the regression analyses. The results were quite similar across mean number correct and number of lapses. Together, the regressors accounted for 20% of the variance in mean number correct and 20% of the variance in number of lapses. Only pretrial pupil variability had a significant direct effect on task performance. Thus, one of the major independent sources of variance in effective WM performance is the degree to which people experience variability in arousal across time.

Table 3 Correlations

	1	2	3	4	5	
1. Mean number correct	–					
2. Number of lapses	– .90***	–				
3. Mean pretrial pupil diameter	.08	– .17	–			
4. Pretrial pupil diameter CoV	– .38***	.37***	– .04	–		
5. Mean task-evoked pupillary response	.26**	– .25*	.07	– .23*	–	
6. Task-evoked pupillary response CoV	– .24*	.20*	.17	.16	– .44***	–

*p < 0.05, **p < 0.01, ***p < 0.001. CoV = coefficient of variation

Table 4 Regression on mean number correct

Predictor	*b*	*t*(102)	*p*
Intercept	3.69	18.76	< .001
Mean pretrial pupil	0.05	0.85	.395
Pretrial pupil CoV	− 11.78	− 3.62	< .001
Mean task-evoked response	0.07	1.10	.274
Task-evoked response CoV	− 0.03	− 1.43	.155

CoV = coefficient of variation

General discussion

The present set of experiments was designed to examine pupillary and subjective correlates of fluctuations in WM performance. The results consistently demonstrated reductions in sustained task-evoked pupil dilations when participants held relatively few items in WM, whereas when they held a relatively large amount of information, their pupil dilated to a greater extent during the delay interval. These results demonstrate that pupil diameter can be used as a reliable index of the attention deployed toward maintaining information in WM. We have previously demonstrated that this is the case across a number of visual WM tasks (Unsworth & Robison, 2015, 2018b). However, the present set of experiments extends this work in two ways. First, in our prior work, the tasks always varied set sizes. In the present studies, the task always gave participants six items to remember. Therefore, any differences in pupil diameter across trials thus reflect how much information people were actually holding in mind, as indicated by their subsequent accuracy at test. The present set of experiments also utilized a whole-report procedure (Adam et al., 2015, 2018; Adam & Vogel, 2016, 2017), rather than a change-detection procedure (Unsworth & Robison, 2015, 2018b). Thus we could estimate the actual quantity of information in WM on individual trials, rather than inferring average performance from a binary report across a number of trials.

Table 5 Regression on number of lapses

Predictor	*b*	*t*(102)	*p*
Intercept	0.15	0.05	.963
Mean pretrial pupil	− 1.74	− 1.85	.067
Pretrial pupil CoV	190.42	3.52	.001
Mean task-evoked response	− 1.04	− 0.99	.323
Task-evoked response CoV	0.46	1.35	.181

CoV = coefficient of variation

The second extension included the use of thought probes to gauge momentary task focus. An assumption of the "lapse" measure in whole-report tasks is that participants temporarily fall into off-task attentional states which cause such lapses in performance (Adam et al., 2015; Adam & Vogel, 2016). To examine whether this is truly the case, we included self-reports of attentional state. Reports of off-task attentional states did indeed follow relatively worse trial performance. In other words, during periods when people reported mind-wandering, external distraction, or absentmindedness, their WM performance suffered. In Experiment 3, we delivered thought probes immediately following lapse trials to try to catch participants in off-task states. Indeed on nearly half of lapse trials, participants reported being in an off-task attention state. So much of the time, such behavioral lapses indicate attentional lapses as well, but half of the time, behavioral lapses were not accompanied by off-task reports, leaving open what other cognitive processes might produce such lapses.

We did not find evidence for differences in pretrial pupil diameter preceding relatively good and relatively poor trials. In one prior study, we observed smaller pretrial pupil diameter preceding errors on low set sizes in a change-detection task (Unsworth & Robison, 2015). We presumed that these small pretrial pupil diameters indicated states of low arousal, which led to lapses in WM. However, we did not observe smaller pretrial pupil diameter preceding lapse trials in the present study. In Experiment 1, the effect was rather large (d = 0.59), with poor trials having larger pre-trial pupils than good trials, in Experiment 2 the effect was small (d = 0.33) with good trials having larger pre-trial pupils than poor trials, and in Experiment 3, the effect was again small (d = -0.23) with poor trials having larger pre-trial pupils than good trials. Thus, the results for pre-trial pupil were quite inconsistent across studies. It is possible that this effect is real and small, and that we did not have enough power in the present design to adequately detect it. However, since the effect was not consistently small and consistently in one direction, this is unlikely. Additionally, in prior work, pretrial pupil diameter does not always show a consistent pattern preceding lapses of attention. In some prior studies, pretrial pupil is significantly smaller preceding reports of mind-wandering and mind-blanking, suggesting people were in a low-arousal state in these instances (Unsworth & Robison, 2016a; Unsworth et al., 2018). In other studies, pretrial pupil diameter is significantly larger preceding reports of mind-wandering (Franklin et al., 2013; Pelagatti et al., 2018). In still other studies, pretrial pupil diameter is about the same preceding on-task and mind-wandering reports (Unsworth & Robison, 2018a). Thus, lapses may not always occur during moments of low arousal, and people can be disengaged from a task in various levels of arousal (Lenartowicz et al., 2013).

We consistently observed differences in task-evoked pupillary responses. These data are actually quite consistent with what Adam et al. (2018) observed with EEG measures during a similar discrete whole-report task. While Adam et al. observed significantly lower CDA amplitude during the delay interval for poor trials compared to good trials, they did not observe a difference in the N2PC ERP, which presumably measures selection of information, nor the P1 component, which presumably measures allocation of spatial attention. However, Adam et al. (2018) did observe reductions in pretrial theta power both preceding and throughout relatively poor trials. An interesting area for future research will be to determine whether pretrial pupil diameter, ERP components during stimulus encoding, CDA, task-evoked pupil diameter, and theta power pick up overlapping or distinct variance in attentional processes in the service of WM.

At the level of individual differences, the results largely corroborate prior investigations of WM capacity and pupillometry (Unsworth & Robison, 2015, 2018b). People who experienced more fluctuations in arousal performed worse on the task, and people who allocated more attention toward maintaining the items during the delay intervals performed better on the task. These findings are in line with prior theorizing about individual differences in the functioning of the LC-NE system in the service of attention and WM (Unsworth & Robison, 2017a, b). We have argued that stability and consistency of delivery of NE to the frontal cortices by the LC is an important underlying factor in why people vary in their ability to control their attention and effectively utilize their limited WM system. When NE is not consistently delivered, arousal tends to fluctuate, which can lead to lapses of attention and be harmful for the performance of tasks that require consistent attention. In the present study, variability in pretrial pupil diameter was the strongest independent predictor of individual differences in task performance, providing further evidence for this account.

Acknowledgements We would like to thank Steven Karmann and Ashley Miller for their assistance in data collection.

References

Adam, K. C. S., Mance, I., Fukuda, K., & Vogel, E. K. (2015). The contribution of attentional lapses to individual differences in visual working memory capacity. *Journal of Cognitive Neuroscience*, *27*, 1601–1616. https://doi.org/10.1162/jocn_a_00811

Adam, K. C. S., Robison, M. K., & Vogel, E. K. (2018). Contralateral delay activity tracks fluctuations in working memory performance. *Journal of Cognitive Neuroscience*, *30*, 1229–1240. https://doi.org/10.1162/jocn_a_01233

Adam, K. C. S., & Vogel, E. K. (2016). Reducing failures of working memory with performance feedback. *Psychonomic Bulletin & Review*, *23*, 1520–1527. https://doi.org/10.3758/s13423-016-1019-4

Adam, K. C. S., & Vogel, E. K. (2017). Confident failures: Lapses of working memory reveal a metacognitive blind spot. *Attention, Perception, & Psychophysics*, *79*, 1506–1523. https://doi.org/10.3758/s13414-017-1331-8

Alnæs, 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*, 1–1. https://doi.org/10.1167/14.4.1

Ariel, R., & Castel, A. D. (2014). Eyes wide open: Enhanced pupil dilation when selectively studying important information. *Experimental Brain Research*, *232*, 337–344. https://doi.org/10.1007/s00221-013-3744-5

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

Aust, F., & Barth, M. (2018). papaja: Create APA manuscripts with R Markdown. Retrieved from https://github.com/crsh/papaja.

Cowan, N. (2001). Metatheory of storage capacity limits. *Behavioral and Brain Sciences*, *24*, 154–176.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, *128*, 309.

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

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*, 252–269. https://doi.org/10.3758/CABN.10.2.252

Heitz, R. P., Schrock, J. C., Payne, T. W., & Engle, R. W. (2008). Effects of incentive on working memory capacity: Behavioral and pupillometric data. *Psychophysiology*, *45*, 119–129. https://doi.org/10.1111/j.1469-8986.2007.00605.x

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

Lenartowicz, A., Simpson, G. V., & Cohen, M. S. (2013). Perspective: Causes and functional significance of temporal variations in attention control. *Frontiers in Human Neuroscience*, *7*, 381. https://doi.org/10.3389/fnhum.2013.00381

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279.

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*, 1532–1543. https://doi.org/10.1111/j.1469-8986.2011.01226.x

Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2012). Memory strength and specificity revealed by pupillometry. *International Journal of Psychophysiology*, *83*, 56–64. https://doi.org/10.1016/j.ijpsycho.2011.10.002

Pelagatti, C., Binda, P., & Vannucci, M. (2018). Tracking the dynamics of mind wandering: Insights from pupillometry. *Journal of Cognition*, *1*, 1–12. https://doi.org/10.5334/joc.41

R Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General*, *126*, 211.

Sara, S. J. (2009). The locus coeruleus and noradrenergic modulation of cognition. *Nature Reviews Neuroscience*, *10*, 211. https://doi.org/10.1038/nrn2573

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2014). Working memory and fluid intelligence: Capacity, attention control, and secondary memory retrieval. *Cognitive Psychology*, *71*, 1–26. Retrieved from https://doi.org/10.1016/j.cogpsych.2014.01.003.

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2015). Working memory delay activity predicts individual differences in cognitive abilities. *Journal of Cognitive Neuroscience*, *27*, 853–865. https://doi.org/10.1162/jocn_a_00765

Unsworth, N., & Robison, M. K. (2015). Individual differences in the allocation of attention to items in working memory: Evidence from pupillometry. *Psychonomic Bulletin & Review*, *22*, 757–765. https://doi.org/10.3758/s13423-014-0747-6

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

Unsworth, N., & Robison, M. K. (2016b). The influence of lapses of attention on working memory capacity. *Memory & Cognition*, *44*, 188–196. https://doi.org/10.3758/s13421-015-0560-0

Unsworth, N., & Robison, M. K. (2017a). A locus coeruleus-norepinephrine account of individual differences in working memory capacity and attention control. *Psychonomic Bulletin & Review*, *24*, 1282–1311. https://doi.org/10.3758/s13423-016-1220-5

Unsworth, N., & Robison, M. K. (2017b). The importance of arousal for variation in working memory capacity and attention control: A latent variable pupillometry study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1962–1987. https://doi.org/10.1037/xlm0000421

Unsworth, N., & Robison, M. K. (2018a). Tracking arousal state and mind wandering with pupillometry. *Cognitive, Affective, & Behavioral Neuroscience*, *18*, 638–664. https://doi.org/10.3758/s13415-018-0594-4

Unsworth, N., & Robison, M. K. (2018b). Tracking working memory maintenance with pupillometry. *Attention, Perception, & Psychophysics*, *80*, 461–484. https://doi.org/10.3758/s13414-017-1455-x

Unsworth, N., Robison, M. K., & Miller, A. L. (2018). Pupillary correlates of fluctuations in sustained attention. *Journal of Cognitive Neuroscience*, *30*, 1241–1253. https://doi.org/10.1162/jocn_a_01251

Unsworth, N., & Spillers, G. J. (2010). Working memory capacity: Attention control, secondary memory, or both? A direct test of the dual-component model. *Journal of Memory and Language*, *62*, 392–406. https://doi.org/10.1016/j.jml.2010.02.001

Vogel, E. K., McCollough, A. W., & Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature*, *438*, 500.