Characterising monitoring processes in event-based prospective memory: Evidence from pupillometry

Joseph Moyes, Nadia Sari-Sarraf, and Sam J. Gilbert

Institute of Cognitive Neuroscience, University College London, UK

Keywords: prospective memory, monitoring, pupillometry

Acknowledgements:
SJG was supported by a Royal Society University Research Fellowship.

Address correspondence to:

Sam Gilbert
Institute of Cognitive Neuroscience
17 Queen Square
London WC1N 3AR
UK
Email: sam.gilbert@ucl.ac.uk
Tel: +44 (0)20 7679 1121
Fax: +44 (0)20 7813 2835

In event-based prospective memory (PM) paradigms, participants are engaged in an ongoing task (e.g. lexical decision) while maintaining an intention to produce a special response if they encounter pre-defined targets (e.g. animal words). This leads to slowed response times even on nontarget trials, which might be caused by: A) a periodic or intermittent process that occurs transiently to check whether the current stimulus is a target, and/or B) a sustained monitoring process maintained throughout task performance rather than being time-locked to stimulus presentation. These processes are hard to distinguish, seeing as the key difference between them occurs in the gap between trials. Processes occurring in these gaps cannot be measured directly by behavioural methods. Here we measured pupil size as a continuous index of intention-related processing in an event-based prospective memory task. Participants performed a lexical decision task while remembering intentions based on either specific target words or categories (e.g. animal words). In two experiments, response times were slowed during PM conditions. Pupil size was significantly increased in the category but not the specific-word condition. This effect was sustained throughout task performance rather than occurring transiently when stimuli were presented. Therefore there was no evidence for a transient pupillometric response associated with nontarget checking, although there was a strong transient response when targets were presented in either PM condition. These results provide evidence for a sustained PM monitoring process that occurrs even in the gaps between trials.

1. Introduction

Event-based prospective memory (EBPM) refers to the ability to remember an intended action when presented with an appropriate cue or event, for example remembering to buy medicine next time you pass a pharmacy. Given the importance of this sort of memory for behavioural independence, EBPM has received increasing attention over the past 30 years (e.g. Brandimonte et al., 1996; Cohen and Hicks, 2017; Kliegel et al., 2008; Kvavilashvili, 1987; McDaniel and Einstein, 2007; Meacham and Leiman, 1982). Much of this research has focused on characterising the cognitive processes underlying this ability. To do so, researchers have developed laboratory tasks that mimic some of the characteristics of everyday EBPM. For example, in a standard laboratory paradigm (e.g. Marsh et al., 2003) participants might be engaged in an ongoing lexical decision task, where they classify a sequence of letter strings as words or nonwords. In prospective memory (PM) conditions, participants are given the additional task of remembering to press a special button if they see an animal word such as 'dog'.

Paradigms such as this can be informative about the processes underlying EBPM in two main respects. First, it is possible to examine PM target trials and investigate factors that influence the likelihood of participants making a PM target response. For example, participants are more likely to remember PM intentions if they encode the intention and encounter the PM cue in the same rather than different rooms (McDaniel et al., 1998). A second way in which experimental paradigms can be informative about EBPM is by examining the influence of PM demands on behaviour on non-target trials where participants simply perform the ongoing task such as lexical decision. A key finding from recent studies has been that response times to nontarget trials are often slower when participants are also tasked with remembering an event-based intention, compared with when they simply perform the ongoing task by itself, even though in both

cases they are presented with the same stimuli and perform the same task (e.g. deciding whether the letter string is a word or nonword). This is sometimes known as the "PM cost" or "PM interference effect" (e.g. Einstein et al., 2005; Heathcote et al., 2015; Smith, 2003).

The existence of the PM cost suggests that some extra cognitive process is occurring in conditions where a PM target is expected, compared with performance of the ongoing task alone. Furthermore, individual differences in the PM cost sometimes correlate with PM accuracy (Smith, 2003), suggesting that the process indexed by this cost may play a causal role in supporting PM success (however this pattern is not always seen and probably only emerges in specific situations; see McNerney and West, 2007). As a result, much debate in recent years has focused on the PM cost and what it tells us about the mechanisms underlying prospective remembering. Insofar as we can characterise the properties of the PM cost, this may provide insight into the process(es) that support EBPM.

Research into the PM cost has often focused on the circumstances under which this behavioural effect is or is not seen. Several studies suggest that when PM responses occur relatively automatically via associative retrieval processes, the PM interference cost is reduced or absent (Einstein et al., 2005; Knight et al., 2011; Scullin et al., 2010b, 2010a). This suggests that the PM cost indexes a monitoring process which is required for PM retrieval whenever automatic processes are insufficient. It has been hypothesised that a key factor determining the need for monitoring is whether the PM task is 'focal' or 'nonfocal'. In a focal task the PM-defining feature is already processed as part of the ongoing task; this is not the case for a nonfocal task. For example, with a lexical decision ongoing task, an instruction to make a PM response if one encounters the word 'dog' would be focal, because the lexical decision task already requires processing of word identity (i.e. to check whether it is a word or not). However, an instruction to make a PM

response if one encounters any animal word would be nonfocal, because the lexical decision task does not require participants to perform semantic categorisation of each stimulus. Studies comparing focal with nonfocal tasks consistently find greater PM costs for nonfocal tasks (Einstein et al., 2005; Harrison et al., 2014; Mullet et al., 2013; Scullin et al., 2010b).

Evidence like this provides support for a 'multiprocess' theory of EBPM (Einstein et al., 2005; McDaniel and Einstein, 2000). This account posits that under some circumstances the presentation of an external cue can lead to spontaneous retrieval of a PM intention, without any PM cost on ongoing trials. The multiprocess account posits that in order to observe cost-free spontaneous PM, it is necessary that a focal task is used (but not sufficient; see Scullin et al., 2010a for discussion). In nonfocal tasks, spontaneous retrieval is inadequate and an additional top-down monitoring process is required, generating PM costs. Therefore nonfocal tasks will always be associated with PM costs, but under certain circumstances focal tasks can be performed without any significant costs.

A large body of evidence has accumulated over the past 15 years to support the multiprocess account, with recent work emphasising a dynamic interplay between monitoring and spontaneous retrieval (Scullin et al., 2013; Shelton and Scullin, 2017). However, it is underspecified at a process level. In particular, the concept of 'monitoring' is not well understood. Similarly, the 'preparatory attentional and memory processes', suggested by Smith (2003) to underlie the PM cost, are not well understood computationally (though see Smith and Bayen, 2004 for a step in this direction). Smith et al. (2014) suggest that the PM cost "is thought to reflect a reallocation of conscious capacity away from the ongoing task in service of processing related to the PM task" (p. 215). However, terms such as "monitoring", "preparatory attention", and "conscious capacity" are hard to map onto well-defined computational processes, limiting their

explanatory value when it comes to phenomena such as the PM cost. In this respect, they play a similar role to the concept of "resource" when it comes to explaining performance costs in dual-task performance (cf Navon, 1984). There is a risk of circularity in their definition. How do we explain slowed RTs in PM conditions? The consumption of conscious capacity. How do we know conscious capacity was consumed? Because RTs were slowed. Unless these theoretical terms are described more precisely, there is a danger that they do little more than redescribe the empirical phenomena under investigation, rather than explaining them.

Progress in clarifying the processes underlying the PM cost has come from Guynn (2003, 2008), who makes an important distinction between sustained and transient monitoring processes. The sustained process is designated "retrieval mode", in analogy with a related concept from the retrospective memory literature (Tulving, 1983). In the context of PM, retrieval mode is conceptualised as "a task set to treat stimuli as cues to retrieve stored intentions" (Guynn, 2008, p. 57). This is "a more or less continuous or constant process that operates after a prospective memory task has been assigned and until it has been completed or cancelled" (Guynn, 2008, p. 57). The transient process is designated "checking", which is a more "periodic or intermittent process" (Guynn, 2008, p. 57) of determining whether stimuli in the current environment constitute PM targets. For example, in the context of a PM task embedded within ongoing lexical decisions, retrieval mode would refer to a sustained readiness to treat each incoming stimulus as a possible PM cue, whereas checking would refer to a transient process that takes place to check whether the current stimulus fits the PM target criteria.

Despite the clear conceptual distinction between these two forms of monitoring process, they are hard to distinguish empirically. Given that the reaction time on any particular trial might be influenced both by any putative sustained monitoring process

and also by any item-specific checking that has occurred on that trial, it is hard to unambiguously identify a behavioural signature of one or the other process. Both processes might contribute to the single response which is observed on a particular trial. Nevertheless, several lines of evidence speak to the distinction between sustained monitoring versus transient checking in EBPM. One approach has been to manipulate experimental factors which are expected to selectively influence retrieval mode versus checking, and investigate their impact on the PM cost. For example, Guynn (2003) compared a situation where experimental and control trials either alternated or were blocked. It was hypothesised that retrieval mode can be engaged or disengaged for a whole block of trials, but not on a trial-by-trial basis. Therefore retrieval mode would apply to both experimental and control trials in the alternating condition, but only experimental trials when they were presented in separate blocks. By contrast, target checking would apply only to experimental trials regardless of condition. Results suggested that both retrieval mode and item-checking were associated with a RT cost.

Another behavioural approach is to consider response-time distributions, rather than just their means. A sustained monitoring process might be expected to generate a general slowing across all trials, whereas a periodic checking process occurring on a subset of trials might lead to increased variance (with positive skew, due to occasional slow trials). Recent studies (Ball and Brewer, 2018; Loft et al., 2014) suggest that PM monitoring may be associated with both general slowing and increased variance, with PM accuracy particularly linked to the general slowing effect.

Although evidence from behavioural studies such as these suggests the existence of both sustained and transient monitoring processes in EBPM, the evidence is necessarily indirect. This is because the key characteristic of a sustained monitoring process is that it occurs not only in an item-specific manner when a stimulus is presented and a response is made, but also in the gap between trials. By definition, behavioural

measures alone are not directly sensitive to processes occurring at this time. Therefore an alternative approach for distinguishing sustained monitoring versus transient checking is to observe neurophysiological measures that can be detected even in the absence of overt behaviour.

One study taking such an approach was conducted by West et al. (2011). This ERP study detected slow-wave PM-related signals up to 1500ms after stimulus presentation, suggesting an electrophysiological correlate of prospective retrieval mode. However, this study did not directly compare putative ERP signatures of retrieval mode versus checking. Another study by Czernochowski et al. (2012) found evidence for a sustained ERP modulation in PM conditions up to 900ms after stimulus presentation, consistent with a sustained monitoring process. This ERP modulation did not differ significantly between two conditions in which PM targets were frequent versus rare, despite large behavioural effects of this manipulation. These results suggest that a similar process of "prospective retrieval mode" may contribute to PM monitoring regardless of target frequency. Czernochowski et al. also found that PM conditions were associated with a larger P2 amplitude 160-210ms after stimulus presentation. This could potentially relate to a target-checking process, however the authors note that it could reflect "increased attention to specific stimulus aspects that are relevant for PM target identification" (p. 74) rather than checking per-se. Furthermore, the key signature of a transient checking process would be a condition x time-bin interaction, i.e. an effect that distinguishes PM from ongoing-only conditions in a temporally-specific or transient manner. However, this interaction effect was not directly investigated by Czernochowski et al.

An alternative approach was taken by Reynolds et al. (2009), who used a mixed blocked and event-related fMRI approach to distinguish sustained versus transient brain activity associated with a PM paradigm. This study revealed brain regions showing both

temporal profiles. However, given the slow haemodynamic response measured by fMRI, Reynolds et al.'s study required the timing of trials to be jittered and unpredictable in order to distinguish transient from sustained effects. Thus, the transient processes identified in their study might occur when stimuli appear unexpectedly (i.e. at unpredictable times, akin to an oddball effect) but it is unclear whether they are also representative of the processes that occur with a predictable periodic stream of stimuli, as in standard PM paradigms (see Cona et al., 2015 for further discussion of transient versus sustained processes in fMRI studies of EBPM).

Here, we aim to add to this debate by using pupillometry as a measure of cognitive load in an EBPM paradigm. Pupillometry involves measurement of participants' pupil diameter as they undergo an experimental task. Even under constant lighting and fixation conditions, pupil diameter reliably increases in response to experimental factors linked to an increase in cognitive demand. For example, Hess and Polt (1964) demonstrated increased pupil diameter associated with mental arithmetic difficulty and Kahneman and Beatty (1966) demonstrated pupil diameter increases time-locked to items added to a digit string maintained in working memory. Recent studies have extended these findings, showing that pupil diameter is also linked to cognitive load in a variety of memory and cognitive control paradigms (see van der Wel and van Steenbergen, 2018, for a recent review). The effect whereby pupil diameter changes in response to a stimulus over the course of an experimental trial is referred to as a task-evoked pupillary response (TEPR). TEPRs begin approximately 0.4 seconds after stimulus onset, peak after around 1 second, and return to baseline around 2-3 seconds after stimulus onset (Goldinger and Papesh, 2012).

Despite the wide variety of cognitive processes investigated with pupillometric methods, to our knowledge this technique has not yet been used to investigate PM (though see West et al., 2007 for a study using eye tracking). Thus, in the present study

we administered a lexical decision task under three conditions: 1) ongoing only, where participants simply performed the ongoing task by itself; 2) single-item PM, where participants maintained an intention to press a special key if they detected a particular word; 3) category PM, where participants maintained an intention to press a special key if they detected a word belonging to a particular semantic category. The latter two conditions have been described as focal and non-focal PM tasks respectively (e.g. Cona et al., 2014). However, the correspondence may not be perfect. For example, in categorical conditions participants may generate exemplars (Scullin et al., 2018), effectively turning nonfocal into focal tasks. Therefore, we use the theoretically neutral terms 'single-item' and 'category' here. Our purpose in this study was to investigate processes involved in monitoring for PM targets, regardless of whether or not targets are actually presented (akin to the PM cost, which reflects a comparison between nontarget trials performed under different experimental conditions). Therefore we focused on pupillometry data from trials in each block prior to the presentation of any targets. This allowed us to detect processes associated with monitoring for PM targets, unconfounded with processes involved in actually detecting and responding to them. We also used a longer-than-usual inter-trial interval of 3 seconds in this study, so that the TEPR could return to baseline between trials. In this way, we aimed to investigate three questions: 1) is it possible to detect the effect of cognitive load in a EBPM paradigm using the method of pupillometry? 2) If so, does this effect differ between a single-item and a category PM task? 3) Insofar as an effect can be detected, is it time-locked to the presentation of each stimulus (consistent with an item-specific checking process), sustained throughout an entire trial (consistent with sustained monitoring), or both?

The temporal profile of a putative sustained monitoring process is straightforward to specify: it is constant rather than time-varying. However, the temporal profile of a transient checking process is harder to define. As noted by Scullin et al.

(2010b) such a process might conceivably occur either before or after processing for the ongoing task is initiated. It should also be noted that a checking process might occur sporadically rather than on every trial. This implies that a variety of pupillometric signatures could potentially correspond with a putative checking process. In searching for pupillometric evidence for a transient checking process we make the following two assumptions. First, we assume that any pupillometric correlate of transient checking would occur time-locked to stimulus presentation, i.e. would occur at approximately the time of stimulus presentation rather than in the inter-trial interval. This is because we instructed participants to make a PM response instead of an ongoing lexical decision response if a target was presented, not after it. As a result of this, any checking process would have to take place before the ongoing response is made, otherwise it would be too late to play any functional role in task performance. Second, we assume that any transient checking process, even if it does not occur every trial, would have to occur sufficiently often to be detectable in the mean pupil response. In other words, we would not be able to detect any transient checking process that occurred extremely infrequently. We note that it may be possible to conceptualise a checking process in accordance with the original proposal of Guynn (2003, 2008) which does not meet these two assumptions. However, such a process could not have a substantial influence on PM performance in the present paradigm, because it would either take place infrequently, or too late to influence responses. We have illustrated three possible patterns of data in Figure 1, as examples of results that would support three potential theoretical models.

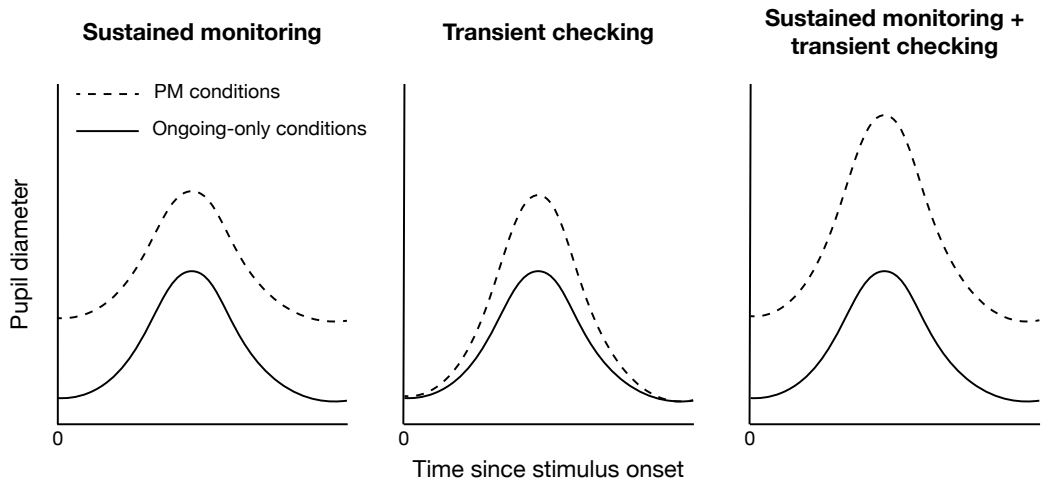

Figure 1. Predictions of three theoretical models. In each case the solid line represents the pupillometric response over the course of a trial in the ongoing-only condition. This is arbitrarily based on a bell-curve shape, as an illustrative example rather than a theoretical prediction. The dotted lines illustrate ways in which the pupillometric response in PM conditions might relate to the ongoing-only response, according to the three models. According to a sustained monitoring model, pupillometric effects should be sustained throughout an entire trial. According to a transient checking model such effects should be time-locked to stimulus presentation. A combination of these two effects is also possible.

2. Experiment 1

2.1 Methods

2.1.1 Participants

Participants were recruited from the Institute of Cognitive Neuroscience subject database, with eligibility contingent on English being their first language to minimise language comprehension effects on lexical decision task performance. 36 participants took part in the study (22 female; age: M=26 years, range 19-57). The study was approved by the UCL Research Ethics Committee (1584/002) and informed written consent was obtained from each participant before taking part.

2.1.2 Design

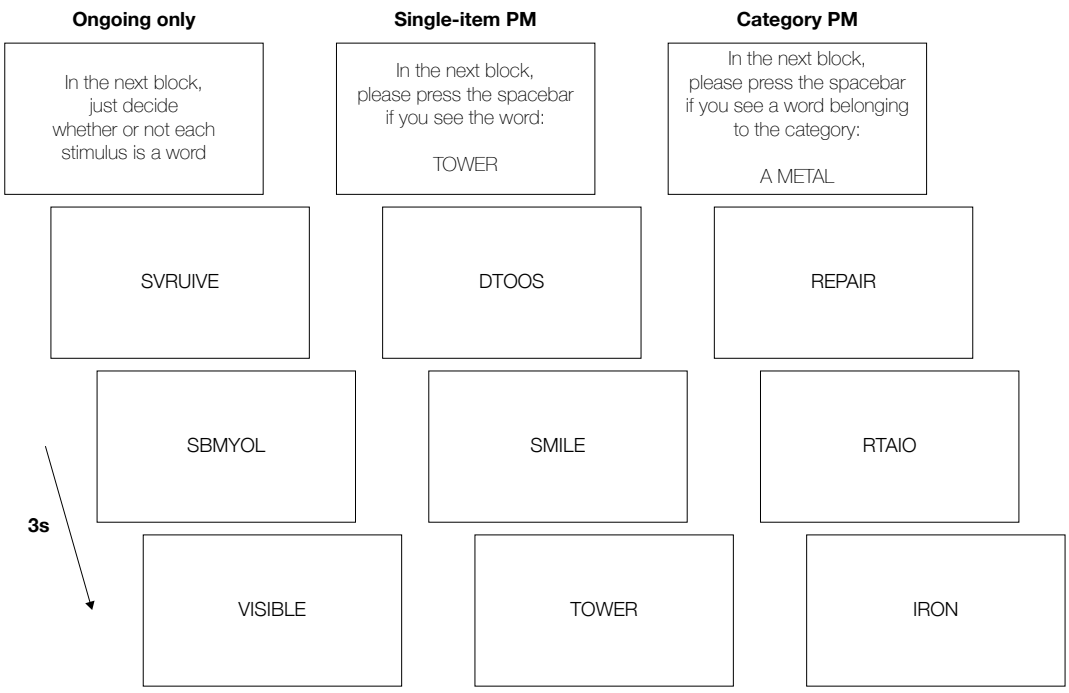

Figure 2. Schematic illustration of the experimental task.

Each participant performed 18 blocks of 30 trials of a lexical decision task, comprising 6 blocks each of the baseline, single-item PM, and category PM conditions (Figure 2; see Cona et al., 2014 for a similar experimental design). Prior to each block, participants performed 10 trials of a pre-block baseline task requiring alternate 'N' and 'M' key presses in response to a neutral stimulus ("XXXXX"), which approximated the word length of lexical decision stimuli. This was included to obtain a baseline pupil diameter reading, which could be subtracted from data from the forthcoming block. The purpose of this was to improve data quality by reducing the impact of fluctuations in pupil size over the course of the experiment. However, preliminary analyses showed that using this correction actually increased noise rather than reducing it. Therefore, we will not consider this baseline correction further, and in the analyses below we simply use the more straightforward measure of raw uncorrected pupillometry data.

Six possible sequences of condition orderings over the 18 blocks were counterbalanced across participants to control for order effects. In each sequence, the conditions appeared a total of six times each and never appeared in two consecutive

blocks. Across the six sequences, each condition was equally likely to follow, and be followed by, each of the other two conditions. This ensured that differences between the conditions could not reflect carryover effects from the previous block.

The frequency and trial positions of PM targets in the single-item and category conditions were matched. Two PM target sequences were generated, assigning PM targets to the 6 blocks of trials for a particular condition. The assignment of these two sequences to the two PM conditions was counterbalanced across participants. In sequence 1, there were two targets in block 1 (12th and 24th trial), one target in block 2 (20th trial), no targets in block 3, two targets in block 4 (22nd and 28th trial), one target in block 5 (26th trial), and no targets in block 6. In sequence 2, there were two targets in block 1 (14th and 20th trial), no targets in block 2, one target in block 3 (22nd trial), no targets in block 4, one target in block 5 (24th trial), and two targets in block 6 (24th and 29th trial). Thus, there was a total of 6 targets in each PM condition.

2.1.3 Equipment and stimuli

Pupil diameter was continuously recorded binocularly at 60 Hz using an EyeTribe eye tracker (http://theeyetribe.com) positioned approximately 60cm from the participant (see Dalmaijer, 2014, for empirical validation of pupillometry data from this device). Participants did not use a chin rest or other immobilisation device. Stimulus presentation and data collection was accomplished with a PC running Psychtoolbox (3.0.12) with MATLAB 8.5, alongside the EyeTribe Toolbox for Matlab (http://github.com/esdalmaijer/EyeTribe-Toolbox-for-Matlab). Stimuli were presented on a 22" monitor approximately 60cm from the participant, running at a refresh rate of 60 Hz, in black 36 point Helvetica font on a dark grey background.

For each category PM block, a word category was selected from the updated version of the category norms developed by Battig and Montague (1969; see Van

Overschelde et al., 2004). From each word category selected, 2 of the top 5 most typical words were selected as potential targets, so that up to two targets in each block could be presented if required. The stimuli for the lexical decision task were generated based on the minimum and maximum HAL (Hyperspace Analogue to Language; Lund and Burgess, 1996) frequency (19,187 and 24,708) and mean word length (4.33 letters) of all 12 category PM target words, using the English Lexicon Project (http://elexicon.wustl.edu/). See Table 1 for a list of all target words and categories.

Block	Single-item PM target	PM category	Possible category-PM targets
1	TOWER	A part of the human body	LEG, ARM
2	GAS	A metal	IRON, STEEL
3	YELLOW	An article of furniture	CHAIR, TABLE
4	BRIDGE	A type of music	JAZZ, POP
5	WINTER	An alcoholic beverage	BEER, WINE
6	SQUARE	A four-footed animal	BEAR, HORSE

Table 1. Single-item and category PM targets in Experiment 1.

The experiment comprised 540 trials in total. Therefore, 270 words (and 270 non-words) matched to the category PM target words were generated. The above matching criteria returned 319 words. Of these, words belonging to the category PM categories were removed, and 6 words were selected as target words for the single-item PM blocks and removed. The remaining total was arbitrarily reduced to 270 words. Each word was used to generate a non-word by randomly flipping two of the letters and checking that the resulting letter string was not an English-language word. The 540 words and non-words were ordered randomly in the experiment. PM target words replaced words/non-words occupying the assigned target trial positions.

2.1.4 Procedure

Participants were tested individually in a dimly lit room with the window blacked out. After calibrating the eye tracker, participants performed a brief practice session consisting of 10 trials of the pre-block baseline task, and 20 trials each of the ongoing-only condition, single-item PM condition, and category PM condition (using different stimuli to those used in the 18 experimental blocks). In the latter two conditions, targets were presented on trials 8 and 16. The experiment began shortly after completion of the practice session.

In each experimental block, including the pre-block baseline task, stimuli were presented for 0.5s, followed by a blank screen for 2.5s before the next trial. Participants could make ongoing or PM responses at any point within this 3s period, but PM responses were only counted as correct if they occurred before an ongoing response. This encouraged participants to engage any periodic checking process at the beginning of each trial, seeing as processes occurring after an ongoing response could not influence accuracy. Using fixed rather than self-paced stimulus presentation parameters ensured that any differences in response time between conditions did not affect visual features of the stimulus display, which could have confounded pupillometry measures. Following each pre-block baseline task there was a 6 second instruction period. In the ongoing-only condition participants were told "In the next block, just decide whether or not each stimulus is a word". In the single-item PM condition participants were told "In the next block, please press the spacebar if you see the word [target word]". In the category PM condition participants were told "In the next block, please press the spacebar if you see any word belonging to the category [target category]". Following the instruction period, the experimental block of 30 trials commenced. Participants pressed the M key to indicate words and the N key to indicate non-words. In the PM conditions, they were asked to press the spacebar instead of the N/M keys to indicate targets. At the end of each block, participants took a brief break until they were ready to start the next block.

2.1.5 Data analysis

Pupil diameter at each timepoint was computed as the mean of the left and right eyes. 180 observations were recorded per trial, and median filtered (order 5) to remove spikes (i.e. the middle timepoint of each 5-frame period was replaced with the median value, however all results were statistically equivalent if this step was omitted). For observations with one eye missing, the other eye was used. For observations with both eyes missing (e.g. blinks), data were linearly interpolated. Only trials before any target was presented were included from each PM block so that any pupillometric difference between conditions reflected prospective monitoring processes on nontarget trials rather than a carryover effect resulting from prior presentation of actual targets. For odd-numbered participants, trial numbers that were excluded from the single-item PM blocks were also excluded from the ongoing-only block (e.g. if trials 14-30 were excluded from the first single-item PM block because a target was presented on trial number 14, the same trials were excluded from the first ongoing-only block). For even-numbered participants, trial exclusions from the category-PM blocks were applied to the ongoing-only blocks. This ensured that at the group level, excluded trials were matched between all three conditions.

Observations were averaged over the trials in each condition and down-sampled to 2Hz for statistical analysis. This pooled the 180 observations (3 seconds x 60Hz) into 6 time bins per trial. This allowed the pupil size data to be analysed in a 3 (Condition) x 6 (Time-bin) repeated measures ANOVA. A significant main effect of Condition would imply a difference in overall pupil size between conditions. A Condition x Time-bin interaction would imply a difference in pupil size between the conditions, which varied in a manner time-locked to each stimulus. This would be consistent with a transient checking effect. By contrast, a sustained monitoring effect would predict a main effect of

Condition but no Condition x Time-bin interaction. Where the assumption of sphericity has not been met (Mauchly's test), Greenhouse-Geisser corrections have been applied. Bayes Factor analyses have been calculated using JASP software (version 0.8.6).

2.2 Results

2.2.1 Behavioural results

Behavioural results are shown in Table 2. PM accuracy did not differ significantly between the single-item and category PM conditions ($F(1,35) = .01$; $p = .91$; $\eta^2_p < .001$), however PM hit responses were made significantly faster in the single-item than the category PM condition ($F(1,35) = 31.4$; $p < .001$; $\eta^2_p = .48$). Furthermore, there was a significant effect of condition on ongoing lexical decision RTs ($F(2,70) = 41.6$; $p < .001$; $\eta^2_p = .54$), and all pairwise comparisons between conditions were significant ($F(1,35) > 10.8$; $p < .003$; $\eta^2_p > .23$). Therefore, a PM cost was observed in both single-item and category PM conditions, and this cost was significantly greater in the category than the single-item PM condition. Ongoing accuracy did not differ significantly between conditions ($F(2,70) = .05$; $p = .95$; $\eta^2_p = .001$). On a small proportion of target trials, participants made an ongoing response followed by a PM response (single-item: M=12.5%, SD=18.0; category: M=14.8%, SD=19.4), suggesting that they detected the target but were unable to make the appropriate response in time. This tendency did not differ between the two conditions ($F(1,35) = .48$; $p = .49$; $\eta^2_p = .01$).

	Ongoing accuracy	Ongoing RT	PM accuracy	PM hit RT
Ongoing only condition	92.2% (9.6)	682 (138)	-	-
Single-item PM condition	92.3% (8.2)	699 (138)	71.3% (27.2)	782 (145)
Category PM condition	92.4% (7.8)	742 (158)	71.8% (29.4)	896 (167)

Table 2. Behavioural results from Experiment 1. Standard deviations are shown in parentheses.

2.2.2 Pupillometry results

Pupillometry results are illustrated in Figure 3 (panel A). There was a significant main effect of Condition on mean pupil size ($F(1.7, 58.9) = 4.0$; $p = .029$; $\eta^2_p = .10$). Pairwise comparisons showed that pupil size was significantly larger in the category PM condition than the ongoing-only condition ($F(1, 35) = 6.3$; $p = .017$; $\eta^2_p = .15$) and marginally significantly larger than the single-item PM condition ($F(1,35) = 3.8$, $p = .059$, $\eta^2_p = .10$). The difference between the baseline and single-item PM condition was not significant ($F(1,35) = .27$; $p = .61$; $\eta^2_p = .01$).

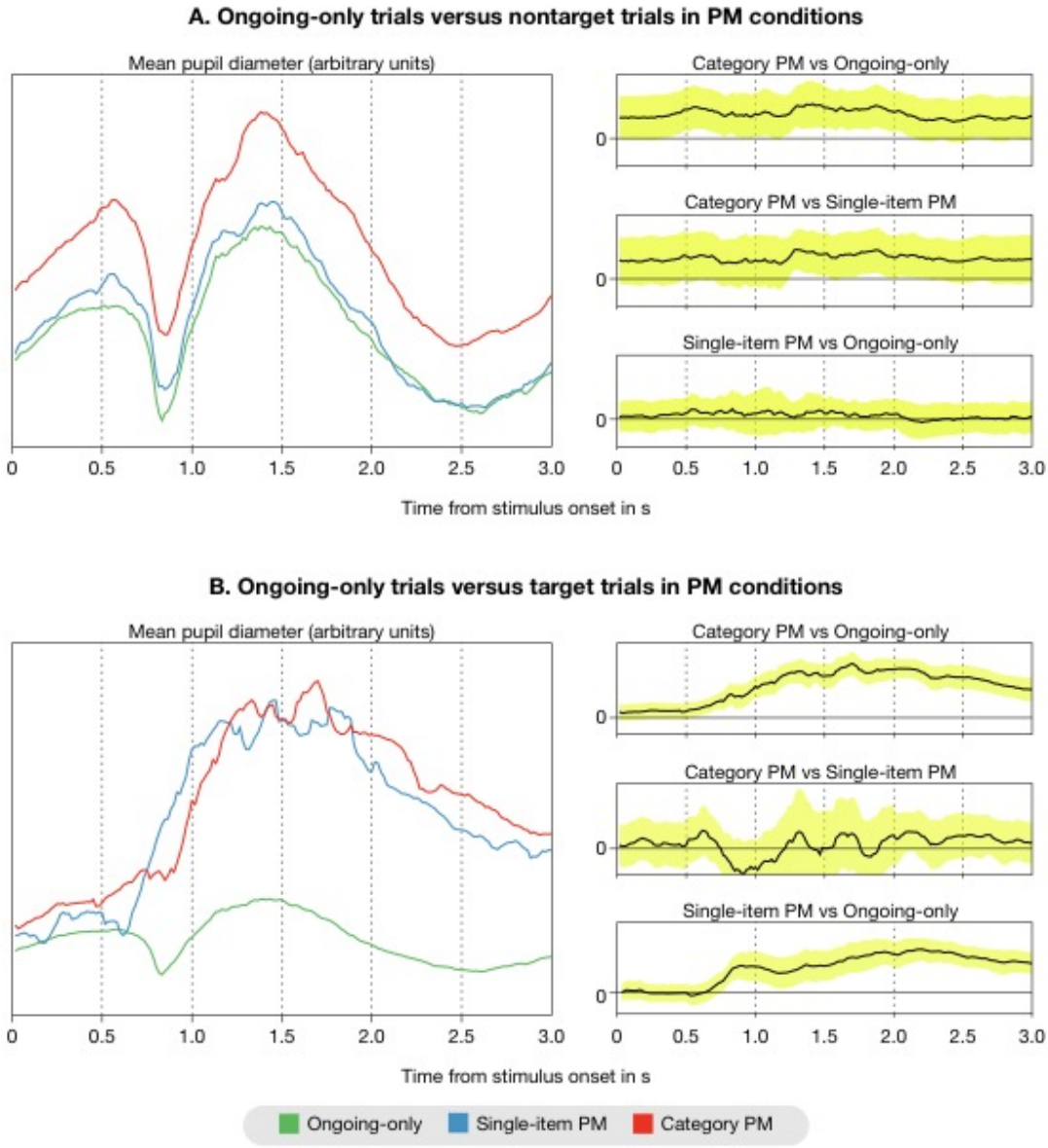

Figure 3. Pupillometry results from Experiment 1 (panel A: nontarget trials; panel B: PM target trials). Graphs on the left show mean pupil size; graphs on the right show pairwise subtractions between conditions. Shaded yellow areas indicate 95% confidence intervals. Therefore there is a significant difference between conditions ($p < .05$) whenever the shaded yellow area does not cross the zero-line.

In addition to the main effect of Condition, there was also a main effect of Time-bin ($F(2.6, 92.3) = 8.9$; $p < .001$; $\eta^2_p = .20$). This shows that there was a significant change in pupil diameter time-locked to the presentation of each stimulus, including a prominent dip at about 0.75s. Previous studies have linked such a dip in pupil size to the effects of visual stimulation (Zénon et al., 2014). However, the Condition x Time-bin interaction was not significant ($F(4.2, 147.6) = .84$; $p = .51$; $\eta^2_p = .02$). Therefore, the effect of PM

condition on pupil diameter appeared to be additive rather than fluctuating in a manner time-locked to the presentation of each stimulus. This pattern is suggestive of sustained monitoring, rather than an item-specific checking process that occurs periodically when stimuli are presented. In order to check whether these results were dependent on choosing 6 time-bins per trial, we conducted equivalent analyses using 3, 9, 15, 18, and 30 time-bins. In every case, results were similar: there was a main effect of Condition ($p < .029$) and of Time-bin ($p < .001$) but no Condition x Time-bin interaction ($p > .21$).

Additional tests showed that the difference between the category PM and ongoing-only conditions was at least marginally significant for all 6 of the time-points considered within each trial, rather than being confined to those time-points close to the presentation of each stimulus ($F(1,35) > 3.8$; $p < .06$; $\eta^2_p > .098$). A similar pattern was found in the comparison between category and single-item PM conditions, where the difference at each time point was at least marginally significant ($F(1,35) > 2.8$; $p < .1$; $\eta^2_p > .076$), with the exception of time point 1 ($F(1,35) = 2.5$, $p = .12$, $\eta^2_p = .07$). Again, this pattern is suggestive of sustained monitoring rather than item-specific checking.

It is of course possible that our failure to observe a significant Condition x Time-bin interaction results from a lack of statistical power to detect such an effect or some other aspect of our design that precludes the detection of time-locked differences between conditions. In order to investigate this possibility, we extracted data from the PM target trials in the category and single-item PM conditions and compared their timecourse with the ongoing trials in the baseline condition (Figure 3, panel B). This showed a highly significant Condition x Time-bin interaction ($F(3.9,137.3) = 12.5$; $p < .001$; $\eta^2_p = .26$). Thus, while the additional processes involved in noticing and responding to actual PM targets yielded a clear pupillometric response that was time-locked to stimulus presentation, there was no such time-locked effect distinguishing the conditions

on nontarget trials. Further analysis showed that both the single-item and category PM conditions showed Condition x Time-bin interactions ($p < .001$) when compared against the ongoing-only condition. In both conditions pupil size was significantly larger than the ongoing-only condition for time bins 2-6 ($p < .004$) but not for bin 1 ($p > .1$). However, the two PM conditions did not differ significantly from each other at any time bin ($p > .3$).

2.2.3 Variability of pupillometry data

Along with mean pupil size in the three conditions, we also analysed variability in the pupillometry data. This allowed us to assess evidence for sporadic monitoring processes. We reasoned that insofar as the single-item and category PM conditions involve monitoring processes that fluctuate over time, this should lead to increased variability in comparison with the ongoing-only condition. We extracted the mean pupil size on a trial-by-trial basis, separately for the three conditions, and calculated the standard deviation of these measures, separately for each participant. We then entered the resulting data into a repeated measures ANOVA. There was a significant effect of condition ($F(2,70) = 4.9$, $p = .01$, $\eta^2_p = .12$), however this reflected *lower* variability in the category PM condition (M=169.7, SD=75.0) than the word PM condition (M = 184.3, SD = 84.2) or the ongoing-only condition (M = 185.6, SD = 92.2). This does not provide evidence for sporadic monitoring in the category PM condition but instead would be more consistent with sporadic processes (e.g. off-task thinking) in the other two conditions.

2.2.4 Bayes factor analysis

The key difference between putative transient checking versus sustained monitoring processes is that the former predicts a Condition x Time-bin interaction but the latter does not. For directly investigating the strength of evidence for the null hypothesis, the Bayes Factor is a more appropriate method than standard frequentist methods. This approach directly compares the likelihood of a null versus an alternative model, given the data (and a pre-specified prior distribution quantifying the likely magnitude of an effect if one were present). We subjected the pupillometry data to a Bayesian repeated measures ANOVA using JASP statistical software with default parameter settings. Compared with a null model including no experimental effects, inclusion of the Condition factor led to a Bayes Factor in favour of the alternative hypothesis of 15233, and inclusion of the Time-bin factor led to a Bayes Factor in favour of the alternative hypothesis of 2×10^{10}. However, additionally including a Condition x Time-bin interaction to the model including both main effects led to a Bayes Factor in favour of the alternative hypothesis of 0.0007. This can be expressed equivalently as a Bayes Factor in favour of the null hypothesis of 1397. Conventionally, Bayes factors in the range 1-3 are considered anecdotal, whereas those in excess of 100 are considered 'extreme' (Jeffreys, 1961). Therefore, while there was extreme evidence for an effect of Condition and Time-bin on pupil size, the evidence for the Condition x Time-bin interaction (predicted by a transient checking model) pointed towards the null. The strength of evidence for this null effect was extreme.

2.3 Discussion

This experiment produced three main results. First, PM condition had a significant effect on pupil size. Second, in comparison with the ongoing-only condition, only category PM led to increased pupil size; single-item PM did not differ significantly from baseline.

Third, the pupil response was sustained throughout each trial, rather than being confined to the period where a stimulus was actually presented and a response made. Therefore in this paradigm, pupillometry provides evidence for the existence of a sustained monitoring process but not transient checking.

However, one limitation of this experiment should be noted when it comes to drawing theoretical conclusions. Although results suggested a difference between category and single-item PM conditions, it is not clear whether this reflects differential monitoring demands imposed by the two conditions, or merely the quantity of information to be remembered. For example, compare a single-item target of 'CASTLE' with a category target of 'A part of the human body'. The former requires a single word to be memorised, but the latter requires participants to memorise a six-word phrase. Therefore, it is not clear whether the difference between conditions reflects different monitoring requirements for single-item versus category-PM conditions, or merely the amount of information to be remembered.

3. Experiment 2

The purpose of this experiment was to replicate the procedure of Experiment 1, but controlling for the mnemonic load of the single-item versus category PM conditions. There were two main differences from Experiment 1. First, the names used for the category PM condition were always single words (e.g. 'Flowers'), which were matched in length and word frequency to the target words used for the single-item PM condition. Second, we removed the pre-block baseline condition used in Experiment 1, seeing as this was not used in the earlier experiment.

3.1 Methods

3.1.1 Participants

As in Experiment 1, 36 participants took part (30 female; age range: 18-49; unfortunately due to data loss we are unable to provide a mean age for this experiment). Participants were drawn from the same participant pool, but none had taken part in the initial experiment.

3.1.2 Materials

The word and nonword stimuli used for the lexical decision task were the same as Experiment 1. However, new stimuli were generated for the PM categories and targets (Table 3) so that the to-be-remembered information presented at the beginning of each block was matched between the single-item and category PM conditions. The category PM categories all consisted of single words. The single-item target words were then matched to these category names in length (category M=6.167; single-item M=6.167) and HAL frequency (category M=12,547; single-item M=12,400; $t(10)=.04$; $p = .97$; $d = .03$).

Block	Single-item PM target	PM category	Possible category-PM targets
1	GARBAGE	Animals	BEAR, HORSE
2	MATCHES	Flowers	ROSE, TULIP
3	HOCKEY	Drinks	COKE, BEER
4	LAWYER	Colors	BLUE, RED
5	DRESS	Fruit	GRAPE, APPLE
6	STEREO	Metals	IRON, STEEL

Table 3. Single-item and category PM targets in Experiment 2.

3.1.3 Procedure

Unlike Experiment 1, there was no pre-block baseline condition in this experiment. The same equipment was used, although in this experiment the pupillometry data was acquired at a rate of 30 Hz rather than 60 Hz (which was still faster than the effective

sampling rate of 2 Hz used in statistical analyses), and participants used a chin rest to minimise head movements. In all other respects, procedures and data analysis methods were the same as Experiment 1. Note that a different testing room was used for this experiment; therefore the lighting conditions were not identical to the earlier experiment but as before a dimly lit room was used with no exposure to natural light.

3.2 Results

3.2.1 Behavioural results

Behavioural results are shown in Table 4. PM accuracy was significantly higher ($F(1,35) = 4.7$; $p = .037$; $\eta^2_p = .12$) and PM hit response times significantly faster ($F(1,35) = 16.6$; $p < .001$; $\eta^2_p = .32$) in the single-item compared with the category PM condition. There was also a significant effect of condition on ongoing lexical decision RTs ($F(2,70) = 29.5$; $p < .001$; $\eta^2_p = .46$) and all pairwise comparisons between conditions were significant ($F(1,35) > 14$; $p < .002$; $\eta^2_p > .28$). Ongoing accuracy did not differ significantly between conditions ($F(2,70) = .41$; $p = .67$; $\eta^2_p = .012$). Therefore results were similar to Experiment 1: both PM conditions incurred a PM cost in terms of slowed RTs to the lexical decision task, and this cost was larger for the category than the single-item PM condition. However, unlike Experiment 1 but consistent with previous research (Marsh et al., 2003), there was also a difference in PM accuracy, which was higher for the single-item than the category PM condition. As in Experiment 1, participants occasionally made an ongoing response followed by a PM response (single-item: M=12.5%, SD=15.6; category: M=14.8%, SD=18.2), suggesting that they detected the target but were unable

to make the appropriate response in time. This tendency did not differ between the two conditions ($F(1,35) = .60$; $p = .44$; $\eta^2_p = .02$).

	Ongoing accuracy	Ongoing RT	PM accuracy	PM hit RT
Ongoing only condition	96.4% (3.6)	657 (120)	-	-
Single-item PM condition	95.9% (5.0)	672 (115)	78.7% (18.1)	910 (211)
Category PM condition	96.1% (3.5)	693 (121)	72.2% (22.2)	1011 (246)

Table 4. Behavioural results from Experiment 2.

3.2.2 Pupillometry results

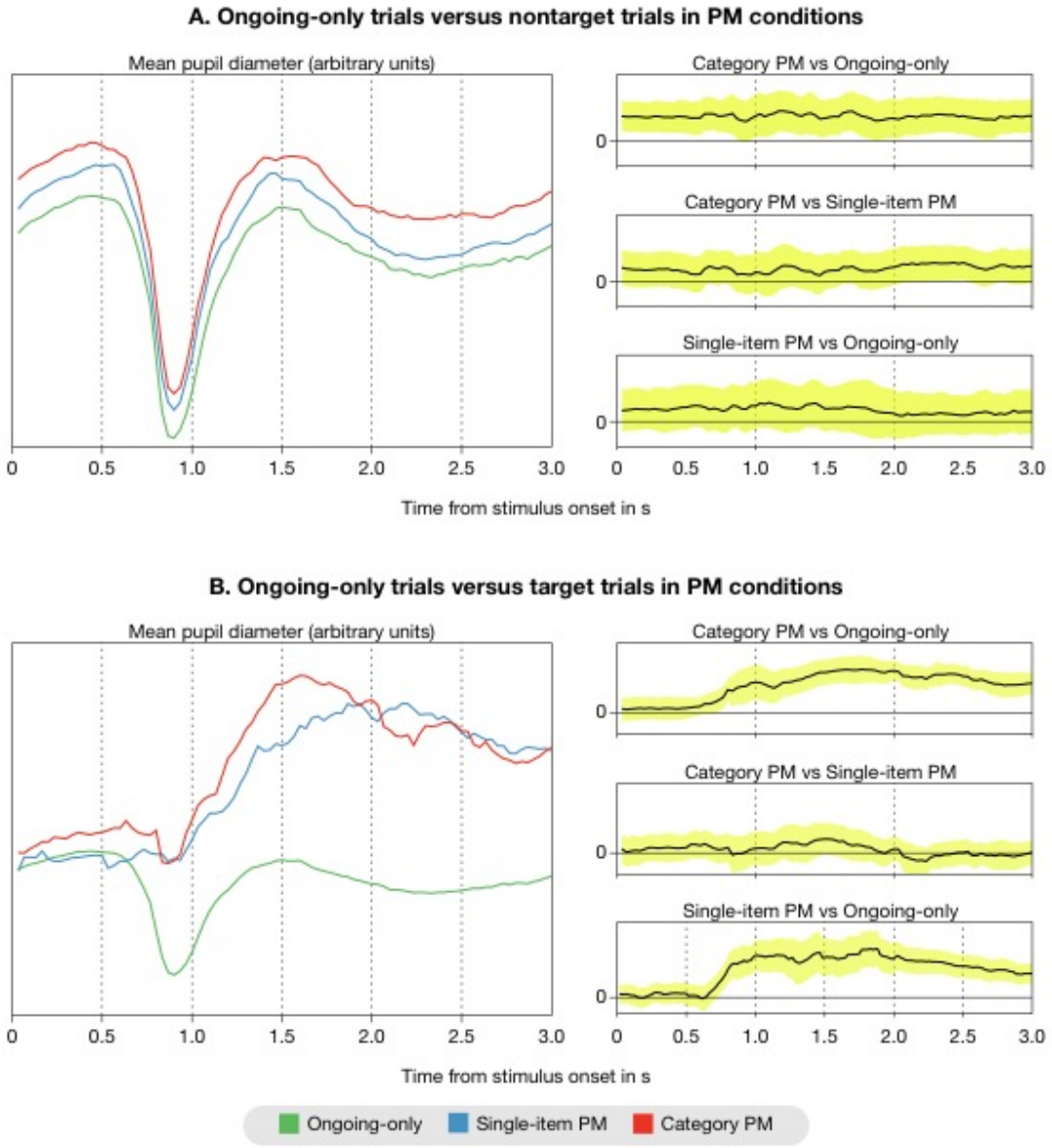

Figure 4. Pupillometry results from Experiment 2 (panel A: nontarget trials; panel B: PM target trials). Graphs on the left show mean pupil size; graphs on the right show pairwise subtractions between conditions. Shaded yellow areas indicate 95% confidence intervals. Therefore there is a significant difference between conditions ($p < .05$) whenever the shaded yellow area does not cross the zero-line.

Results were similar to Experiment 1 (see Figure 4). There was a main effect of Condition on mean pupil size ($F(2,70) = 3.7$; $p = .03$; $\eta^2_p = .10$). Pairwise comparisons showed that pupil size was significantly larger in the category PM condition than the baseline condition ($F(1,35) = 8.8$; $p = .005$; $\eta^2_p = .20$). The comparison between the single-item and category PM conditions was not significant ($F(1,35) = 2.5$; $p = .12$; $\eta^2_p =$

.07); nor was the comparison between baseline and single-item PM conditions ($F(1,35) = 1.2$, $p = .27$, $\eta^2_p = .03$).

There was also a main effect of Time-bin ($F(2.5, 87.8) = 6.1$, $p = .002$; $\eta^2_p = .15$), but the Condition x Time-bin interaction was not significant ($F(6.0, 208.9) = .74$; $p = .62$; $\eta^2_p = .02$). As in Experiment 1, we repeated these analyses using 3, 9, 15, 18, and 30 time-bins per trial. In all analyses there was a significant main effect of Condition ($p < .031$) but no significant Condition x Time-bin interaction ($p > .41$), replicating the 6-bin analysis. The main effect of Time-bin was significant ($p < .001$) in all analyses apart from 3-bin ($p = .34$).

The difference between category PM and baseline conditions was observed at all 6 time points considered within each trial, rather than being confined to the time-points close to the presentation of each stimulus ($F(1,35) > 6.9$; $p < .02$; $\eta^2_p > .16$). The comparison between category and single-item PM conditions was significant for time point 5 ($F(1,35) = 4.4$, $p = .04$, $\eta^2_p = .11$) and marginally significant for time point 6 ($F(1,35) = 3.1$, $p = .09$, $\eta^2_p = .08$); all other time points were nonsignificant ($F(1,35) < 2.1$; $p > .16$; $\eta^2_p < .06$). In sum, results were similar to Experiment 1 in pointing towards a sustained monitoring effect in the category PM condition rather than transient item-specific checking.

Results were also similar to Experiment 1 in showing a highly significant Condition x Time-bin interaction ($F(5.5, 193.9) = 12.2$; $p < .001$; $\eta^2_p = .26$) when target trials were used as data for the two PM conditions rather than nontarget lexical decision trials. Thus, our paradigm was sensitive to a time-locked effect distinguishing PM target trials from ongoing lexical decision trials (despite each participant only receiving six PM target trials per condition), but no time-locked effect distinguishing lexical decision trials between the three experimental conditions. As in Experiment 1, the Condition x Time-

bin interaction was significant for both PM conditions when compared against the ongoing-only condition ($p < .001$). Both conditions differed significantly from the ongoing-only condition in time-bins 2-5 ($p < .02$) but not in bin 1 ($p > .5$). The two PM conditions did not differ significantly each other at any time bin ($p > .2$).

3.2.3 Variability of pupillometry data

As in Experiment 1 we analysed the variability of the pupillometry data in the three conditions. Variability was similar in the three conditions (ongoing-only: M=225.2, SD=95.8; single-item PM: M=232.0, SD=118.7; category PM: M=232.1, SD=114.4) and did not differ significantly ($F(1.7, 59.4) = .83$, $p = .42$, $\eta^2_p = .02$). Therefore, like Experiment 1 there was no evidence for a sporadic monitoring process in the PM conditions, however unlike the earlier experiment nor was there evidence for reduced variability in the category PM condition.

3.2.4 Bayes factor analyses

Compared with a null model including no experimental effects, the Bayes Factor of a model including the Condition factor was 1128 and the Bayes Factor of a model including the Time-bin factor was 4.5×10^7. However, additionally including a Condition x Time-bin interaction to the model including both main effects led to a Bayes Factor of 0.0007. This is equivalent to a Bayes Factor in favour of the null hypothesis of 1463. Therefore as well as extreme evidence for an effect of Condition and Time-bin, there was extreme evidence for a null effect of the Condition x Time-bin interaction.

3.2.5 Cross-experiment comparisons

We investigated consistency of results across the two experiments by entering pupillometry data from nontarget trials into a single ANOVA with within-subject factors of Condition and Time-bin, and a between-subject factor of Experiment. There was a significant main effect of Condition ($F(2, 140) = 7.4$; $p < .001$; $\eta^2_p = .096$). Pairwise comparisons showed a significant difference between the category PM condition and the ongoing-only condition ($F(1,70) = 14.9$; $p < .001$; $\eta^2_p = .176$) and between category PM condition and the single-item PM condition ($F(1,70) = 6.3$; $p = .015$; $\eta^2_p = .082$). However, the ongoing-only and single-item PM conditions did not differ significantly ($F(1,70) = 1.5$; $p = .23$; $\eta^2_p = .02$). There was also a main effect of Time-bin ($F(2.7, 187.1) = 9.5$; $p < .001$; $\eta^2_p = .12$), qualified by a Time-bin x Experiment interaction ($F(2,7, 187.1) = 4.9$; $= .004$; $\eta^2_p = .066$). This could reflect the pupillary response to visual stimulation, modulated by different lighting conditions in the two experiments. No other effects, including the Condition x Time-bin interaction predicted by the transient checking model, were significant ($p > .35$).

3.3 Discussion

Results from Experiment 2 were similar to those from Experiment 1 with the exception that there was a significant difference in PM accuracy between the two PM conditions in this experiment. Thus, results from this study converge with Experiment 1 in showing that differences between performing an ongoing task in PM versus baseline conditions can be detected in pupillometric data. Further, they are consistent with the earlier experiment in showing a significant effect for category but not single-item PM

conditions, and additionally demonstrate that this is not due to the mnemonic load of the two conditions seeing as the instructions were matched in word length and frequency. Finally, as in Experiment 1, this experiment shows that the pupillometric effect distinguishing the conditions occurred in a sustained manner throughout each trial rather than being time-locked to stimulus presentation.

4. Effect of word versus non-word stimuli

In the foregoing analyses of the two experiments we have collapsed over lexical decision trials performed with word versus nonword stimuli, for simplicity. However previous behavioural research has suggested that this might be an important factor for understanding sustained versus transient monitoring processes in EBPM. Cohen et al. (2012) investigated a single-item EBPM task embedded within an ongoing lexical decision task. They found that the PM cost was most pronounced on nontarget trials that matched the category of PM target. That is, when the PM target was a word (e.g. GIRL), the PM cost was most pronounced on ongoing word trials. When the PM target was a nonword (e.g. UEBL) the PM cost was most pronounced on ongoing nonword trials. Cohen et al. (2012) interpret these findings in the context of Guynn's (2003) monitoring theory, explaining the PM cost on trials matching the PM target category in terms of a checking process that operates only when the stimulus is a possible PM target. In a related study, Lourenço and Maylor (2014) found that the PM cost was larger for ongoing trials that matched the PM target category, but a smaller PM cost was also detected on target nonmatch trials.

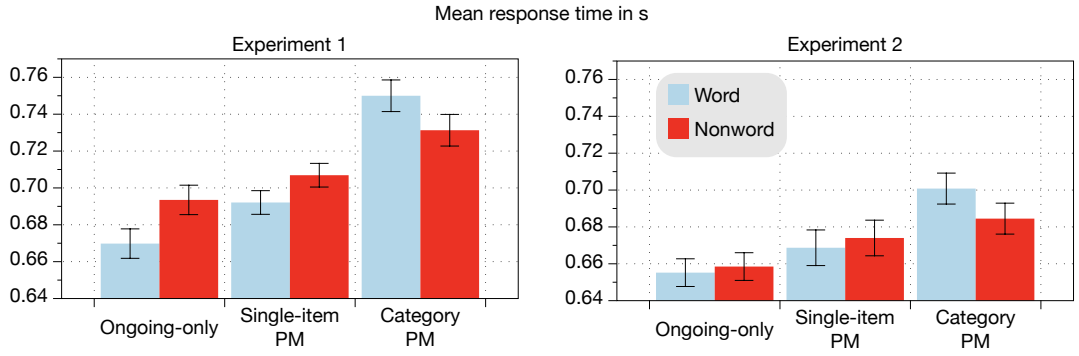

Figure 5. Mean lexical decision response times, presented separately for word versus nonword stimuli. Error bars represent 95% confidence intervals for the within-subject comparison between word vs nonword in each condition, such that a significant effect ($p < .05$) is indicated by nonoverlapping bars.

We re-analysed the behavioural and pupillometry data described above to investigate any effect of word/nonword status. Behavioural results are shown in Figure 5. In each experiment, a significant PM cost was observed for both word and nonword trials, in both single-item and category-PM conditions ($F(1,35) > 5.6$; $p < .03$; $\eta^2_p > .13$). There was also a significant difference between single-item and category PM ongoing RTs for both words and nonwords ($F(1,35) > 11.7$; $p < .002$; $\eta^2_p > .25$), with the exception of nonwords in Experiment 2 which showed a marginally significant effect ($F(1,35) = 3.3$; $p = .079$; $\eta^2_p = .085$). However, ANOVAs investigating Lexicality (word, nonword) x Condition (ongoing, single-item, category) showed a significant interaction between the two factors (Experiment 1: $F(1.6,57.4) = 16.2$; $p < .001$; $\eta^2_p = .32$; Experiment 2: $F(2, 70) = 5.2$; $p = .009$; $\eta^2_p = .13$). In Experiment 1, RTs to words were significantly faster than nonwords in the ongoing and single-item PM conditions, but significantly *slower* in the category-PM condition ($F(1,35) > 4.8$; $p < .04$; $\eta^2_p > .12$). In Experiment 2 all three trends were in the same direction but none of the comparisons were significant ($F(1,35) < 4.0$; $p > .05$; $\eta^2_p < .11$). These results suggest that ongoing word trials in the category PM condition were particularly slowed by a checking process to determine whether the word belonged to the PM target category.

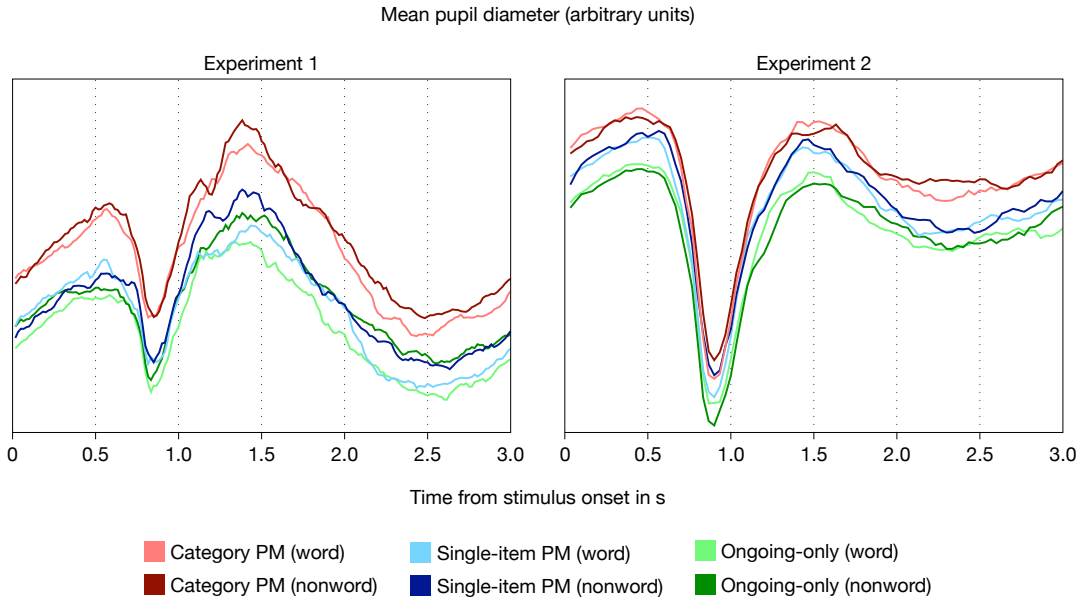

Figure 6. Pupillometry data for nontarget trials, separated for word vs nonword stimuli.

Turning now to the pupillometry data (Figure 6), these results were analysed in Lexicality x Condition x Time-bin ANOVAs. In both experiments the main effect of Condition remained significant (Experiment 1: $F(1.7, 59.1) = 4.1$; $p = .03$; $\eta^2_p = .10$; Experiment 2: $F(2,70) = 3.6$; $p = .03$; $\eta^2_p = .09$) and the Condition x Time-bin interactions remained nonsignificant ($F < 1$). The critical effect which would be suggestive of an item-specific checking process specifically on category PM word trials would be the Lexicality x Condition x Time-bin interaction. This interaction did not approach significance in either experiment (Experiment 1: $F(5.5,191.7) = .70$; $p = .64$; $\eta^2_p = .02$; Experiment 2: $F(5.8, 204.0) = .63$; $p = .70$; $\eta^2_p = .02$). In Experiment 1 there was a significant main effect of Lexicality ($F(1,35) = 6.2$; $p = .02$; $\eta^2_p = .15$), reflecting greater mean pupil size on nonword than word trials. However, this effect was nonsignificant in Experiment 2 ($F(1,35) = .1$; $p = .76$; $\eta^2_p = .003$) and none of the interactions involving the Lexicality factor were significant in either experiment ($p > .09$). In sum, while the behavioural data

were suggestive of both sustained monitoring and item-specific checking processes (consistent with earlier studies, e.g. Guynn, 2003), the pupillometry data revealed evidence for sustained monitoring alone, consistent with the analyses presented above.

5. General Discussion

This study used pupillometry in an attempt to investigate cognitive processes involved in EBPM. There were three main findings: 1) performance of a PM task was associated with significantly increased pupil dilation, compared with performance of an ongoing task alone; 2) this effect was observed for category PM but not single-item PM; and 3) PM-related pupil dilation occurred in a sustained manner rather than being time-locked to stimulus presentation.

These results provide evidence for a sustained monitoring or retrieval mode process in EBPM (Guynn, 2003), at least when targets are defined by categories rather than single items. By definition, behavioural methods are not directly sensitive to processes occurring in the gaps between trials. Our findings provide direct evidence that at least some EBPM conditions are associated with a process that occurs not only when stimuli are presented, but also during the inter-trial interval, as predicted by models of EBPM that posit a sustained monitoring process.

The present results did not provide any evidence for item-specific checking in EBPM. However, behavioural effects did suggest target checking on category PM word trials. Therefore, these results suggest a divergence between the processes detected by RT versus pupillometric measures. Insofar as there are separable sustained and transient processes contributing to EBPM (Guynn, 2003), pupillometry may be particularly sensitive to the former type of process. This is unlikely to reflect the sluggish timecourse of the pupillometric response, seeing as our inter-trial-interval of 3s was sufficient to

allow a return to baseline (Goldinger and Papesh, 2012). Thus, evidence for a sustained process does not only rest on the absence of Condition x Time-bin interactions, but also on the existence of significant pupillometric effects even at timepoints that were temporally remote from stimulus presentation. Our results also cannot be attributed to a lack of statistical power to detect any sort of pupillometric effect time-locked to stimulus presentation, seeing as highly significant Condition x Time-bin interactions were obtained when comparing PM target trials with ongoing lexical condition trials. This effect was seen for target trials in both PM conditions, which did not differ from each other. Therefore, insofar as the single-item and category PM conditions differed in their cue detection processes on target trials (e.g. spontaneous retrieval versus monitoring-based processes), this was not detectable in the pupillometry data.

Of course, the present results do not exclude the possibility that other paradigms might find pupillometric evidence for transient checking in EBPM. It is possible that in our paradigm transient checking was not detectable in the pupillometry data because it only occurred on a small proportion of trials. In this case, further studies that manipulate factors such as the frequency of PM targets might lead to a more readily detectable pupillometric response by affecting the balance between sustained monitoring and transient checking processes (Czernochowski et al., 2012). Furthermore, several features of our paradigm may have particularly encouraged the use of sustained monitoring processes. First, our ITI of 3s was rather long. This may have led participants to use the interval between trials for self-remindings of future intentions (Hicks et al., 2000; Sellen et al., 1997). Second, participants were instructed to make PM responses instead of ongoing responses if a target was presented. This may have increased the cognitive load of the task, encouraging sustained monitoring (however an alternative interpretation would be that this would bias participants towards transient checking upon stimulus presentation, seeing as any process following the production of an ongoing response

would be too late to contribute towards the instructed PM behaviour). Third, the ongoing, single-item PM, and category PM blocks were intermixed, which could have increased monitoring so that participants could keep track of which condition currently applied. Fourth, participants were informed of PM intentions immediately before the relevant block of trials. This contrasts with standard PM paradigms where there is a filled gap between intention encoding and PM task performance, to prevent continuous rehearsal of the intention. Therefore, a full characterisation of the relationship between the task-evoked pupillary response, sustained monitoring, and transient checking requires further investigation of factors such as the ones discussed above.

How might we characterise the sustained PM-related process detected in the current study? One possibility comes from the computational model of EBPM presented by Gilbert et al. (2013). This model contains a 'monitoring' node which can be activated in a sustained manner during task performance. Activation of this node has the effect of boosting the ability of incoming stimuli to drive activity in 'target detection' nodes, which can lead to the production of a PM response. This can be seen as a computational implementation of Guynn's (2003) theorised retrieval mode process.

Another way of characterising the sustained PM-related process is in terms of global parameters which affect behaviour in computational frameworks such as drift diffusion (Ratcliff, 1978) or linear ballistic accumulator (Brown and Heathcote, 2008) models. In particular, it has been suggested that PM conditions are associated with a shift in response threshold in such models. According to 'delay theory', participants strategically adopt a more conservative response threshold in PM conditions (Heathcote et al., 2015; Loft and Remington, 2013; see also Boywitt and Rummel, 2012; Horn and Bayen, 2015 for related work). This means that there is more time for PM-related evidence to accumulate before a potentially-erroneous ongoing response is made.

Threshold shifts can perhaps be seen as a simple instantiation of "retrieval mode" in the sense that allowing more time for PM-related evidence to accumulate can be seen as one mechanisms by which we "treat stimuli as cues to retrieve stored intentions" (Guynn, 2012, p. 57). Recent work using pupillometry has suggested a link specifically between pupil dilation and decision threshold in drift diffusion models (Cavanagh et al., 2014). This implies that increased pupil dilation might not necessarily reflect cognitive effort per se, but could also relate to parameters relating to cautiousness such as decision threshold. This may explain why the pupillometry results showed a more complex pattern than simply mirroring RT effects. It is also in line with PM research linking sustained monitoring with decision threshold in drift diffusion models (Horn & Bayen, 2015). Therefore, the present results are in some respects highly compatible with delay theory.

However, delay theory can encounter difficulties when it comes to explaining the variety of behavioural results associated with distinct types of PM task and behavioural strategies, using a single parameter such as decision threshold (see Anderson, Rummel, & McDaniel, 2018, for discussion). Furthermore, in the present study both single-item and category PM conditions were associated with a RT cost on nontarget trials, but only the category PM condition was associated with a pupillometric effect. It is not clear how delay theory would explain this divergence. This finding fits comfortably, however, with the multiprocess framework (McDaniel and Einstein, 2000), seeing as the two conditions correspond to nonfocal and focal PM respectively. Another problematic aspect of delay theory is to understand how it links with PM behaviour in everyday life. While delay theory may give a good account of how participants behave in laboratory tasks, it seems implausible as an account of real-world PM that individuals simply go about their lives delaying behaviour – all behaviour – as a mechanism for allowing intended actions to be

produced. Thus, this account requires further work to clarify its implications for real-world PM.

An important question for further research will be to investigate the temporal profile and computational signature of the sustained monitoring process detected in the current study. How rapidly can this process be switched on and off, how is it affected by experimental factors such as target frequency, and how far is it under the voluntary control of the participant rather than being directly driven by stimulus input? The present results suggest that pupillometry can be a suitable technique for answering these questions.

References

Anderson, F.T., Rummel, J., McDaniel, M.A., 2018. Proceeding with care for successful prospective memory: Do we delay ongoing responding or actively monitor for cues? J. Exp. Psychol. Lean. Mem. Cogn. 44, 1036-1050. doi:10.1037/xlm0000504

Ball, B.H., Brewer, G.A., 2018. Proactive control processes in event-based prospective memory: Evidence from intraindividual variability and ex-gaussian analyses. J. Exp. Psychol. Learn. Mem. Cogn. 44, 793–811. doi:10.1037/xlm0000489

Battig, W.F., Montague, W.E., 1969. Category norms of verbal items in 56 categories A replication and extension of the Connecticut category norms. J. Exp. Psychol. 80, 1–46. doi:10.1037/h0027577

Boywitt, C.D., Rummel, J., 2012. A diffusion model analysis of task interference effects in prospective memory. Mem. Cognit. 40, 70–82. doi:10.3758/s13421-011-0128-6

Brandimonte, M.A., Einstein, G.O., McDaniel, M.A., 1996. Prospective Memory: Theory and Applications. Psychology Press.

Brown, S.D., Heathcote, A., 2008. The simplest complete model of choice response time: Linear ballistic accumulation. Cogn. Psychol. 57, 153–178. doi:10.1016/j.cogpsych.2007.12.002

Cavanagh, J.F., Wiecki, T. V, Kochar, A., Frank, M.J., 2014. Eye tracking and pupillometry are indicators of dissociable latent decision processes. J. Exp. Psychol. Gen. 143, 1476–1488. doi:10.1037/a0035813

Cohen, A.-L., Hicks, J.L., 2017. Prospective Memory Remembering to Remember , Remembering to Forget.

Cohen, A., Jaudas, A., Hirschhorn, E., Sobin, Y., Peter, M., 2012. The specificity of prospective memory costs The specificity of prospective memory costs 20, 37–41.

Cona, G., Bisiacchi, P.S., Moscovitch, M., 2014. The effects of focal and nonfocal cues

on the neural correlates of prospective memory: Insights from ERPs. Cereb. Cortex. doi:10.1093/cercor/bht116

Cona, G., Scarpazza, C., Sartori, G., Moscovitch, M., Bisiacchi, P.S., 2015. Neural bases of prospective memory: A meta-analysis and the "Attention to Delayed Intention" (AtoDI) model. Neurosci. Biobehav. Rev. 52, 21–37. doi:10.1016/j.neubiorev.2015.02.007

Czernochowski, D., Horn, S., Bayen, U.J., 2012. Does frequency matter? ERP and behavioral correlates of monitoring for rare and frequent prospective memory targets. Neuropsychologia 50, 67–76. doi:10.1016/j.neuropsychologia.2011.10.023

Dalmaijer, E.S., 2014. Is the low-cost EyeTribe eye tracker any good for research? PeerJ Prepr. 4, 1–35. doi:10.7287/peerj.preprints.141v2

Einstein, G.O., McDaniel, M.A., Thomas, R., Mayfield, S., Shank, H., Morrisette, N., Breneiser, J., 2005. Multiple processes in prospective memory retrieval: factors determining monitoring versus spontaneous retrieval. J. Exp. Psychol. Gen. 134, 327–42. doi:10.1037/0096-3445.134.3.327

Gilbert, S.J., Hadjipavlou, N., Raoelison, M., 2013. Automaticity and control in prospective memory: a computational model. PLoS One 8, e59852. doi:10.1371/journal.pone.0059852

Goldinger, S.D., Papesh, M.H., 2012. Pupil Dilation Reflects the Creation and Retrieval of Memories. Curr. Dir. Psychol. Sci. 21, 90–95. doi:10.1177/0963721412436811

Guynn, M.J., 2008. Theory of monitoring in prospective memory: Instantiating a retrieval mode and periodic target checking., in: Prospective Memory: Cognitive, Neuroscience, Developmental, and Applied Perspectives. Erlbaum, New York, pp. 53–72.

Guynn, M.J., 2003. A two-process model of strategic monitoring in event-based prospective memory: Activation/retrieval mode and checking. Int. J. Psychol. 38,

245–256. doi:10.1080/00207590244000205

Harrison, T.L., Mullet, H.G., Whiffen, K.N., Ousterhout, H., Einstein, G.O., 2014. Prospective memory: Effects of divided attention on spontaneous retrieval. Mem. Cognit. 42, 212–224. doi:10.3758/s13421-013-0357-y

Heathcote, A., Loft, S., Remington, R., 2015. Slow Down and Remember to Remember!! A Delay Theory of Prospective Memory Costs. Psychol Rev 122, 376–410. doi:http://dx.doi.org.ez.statsbiblioteket.dk:2048/10.1037/a0038952

Hess, E.H., Polt, J.M., 1964. Pupil Size in Relation to Mental Activity during Simple Problem-Solving. Science 143, 1190–2. doi:10.1126/science.143.3611.1190

Hicks, J.L., Marsh, R.L., Russell, E.J., 2000. The Properties of Retention Intervals and Their Affect on Retaining Prospective Memories. J. Exp. Psychol. Learn. Mem. Cogn. 26, 1160–1169. doi:10.1037/0278-7393.26.5.1160

Horn, S.S., Bayen, U.J., 2015. Modeling criterion shifts and target checking in prospective memory monitoring. J. Exp. Psychol. Learn. Mem. Cogn. 41, 95–117. doi:10.1037/a0037676

Jeffreys, H., 1961. Theory of Probability, Theory of Probability.

Kahneman, D., Beatty, J., 1966. Pupil Diameter and Load on Memory. Science (80-.). 154, 1583–1585. doi:10.1126/science.154.3756.1583

Kliegel, M., McDaniel, M.A., Einstein, G.O., 2008. Prospective memory: Cognitive, neuroscience, developmental, and applied perspectives., Prospective memory: Cognitive, neuroscience, developmental, and applied perspectives. Erlbaum, Mahwah. doi:10.4324/9780203809945

Knight, J.B., Meeks, J.T., Marsh, R.L., Cook, G.I., Brewer, G. a, Hicks, J.L., 2011. An observation on the spontaneous noticing of prospective memory event-based cues. J. Exp. Psychol. Learn. Mem. Cogn. 37, 298–307. doi:10.1037/a0021969

Kvavilashvili, L., 1987. Remembering intention as a distinct form of memory. British 4.

Loft, S., Bowden, V.K., Ball, B.H., Brewer, G. a., 2014. Fitting an ex-Gaussian function to examine costs in event-based prospective memory: Evidence for a continuous monitoring profile. Acta Psychol. (Amst). 152, 177–182. doi:10.1016/j.actpsy.2014.08.010

Loft, S., Remington, R.W., 2013. Wait a second: Brief delays in responding reduce focality effects in event-based prospective memory. Q. J. Exp. Psychol. 66, 1432–1447. doi:10.1080/17470218.2012.750677

Lourenço, J.S., Maylor, E.A., 2014. Is it relevant? Influence of trial manipulations of prospective memory context on task interference. Q. J. Exp. Psychol. 67, 687–702. doi:10.1080/17470218.2013.826257

Lund, K., Burgess, C., 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. Behav. Res. Methods, Instruments, Comput. 28, 203–208. doi:10.3758/BF03204766

Marsh, R.L., Hicks, J.L., Cook, G.I., Hansen, J.S., Pallos, A.L., 2003. Interference to ongoing activities covaries with the characteristics of an event-based intention. J. Exp. Psychol. Learn. Mem. Cogn. 29, 861–70. doi:10.1037/0278-7393.29.5.861

McDaniel, M.A., Einstein, G.O., 2007. Prospective Memory: An Overview and Synthesis of an Emerging Field. Sage Publications Ltd, Los Angeles.

McDaniel, M.A., Einstein, G.O., 2000. Strategic and automatic processes in prospective memory retrieval: a multiprocess framework. Appl. Cogn. Psychol. 14, S127–S144. doi:10.1002/acp.775

McDaniel, M.A., Robinson-Riegler, B., Einstein, G.O., 1998. Prospective remembering: Perceptually driven or conceptually driven processes? Mem. Cogn. 26, 121–134. doi:10.3758/BF03211375

McNerney, M.W., West, R., 2007. An imperfect relationship between prospective memory and the prospective interference effect. Mem. Cognit. 35, 275–82.

Meacham, J.A., Leiman, B., 1982. Remembering to perform future actions, in: Memory Observed: Remembering in Natural Contenxts. pp. 327–336.

Mullet, H.G., Scullin, M.K., Hess, T.J., Scullin, R.B., Arnold, K.M., Einstein, G.O., 2013. Prospective memory and aging: Evidence for preserved spontaneous retrieval with exact but not related cues. Psychol. Aging 28, 910–922. doi:10.1037/a0034347

Navon, D., 1984. Resources--a theoretical soup stone? Psychol. Rev. 91, 216–234. doi:10.1037//0033-295X.91.2.216

Ratcliff, R., 1978. A theory of memory retrieval. Psychol. Rev. 85, 59–108. doi:10.1037/0033-295X.85.2.59

Reynolds, J.R., West, R., Braver, T., 2009. Distinct neural circuits support transient and sustained processes in prospective memory and working memory. Cereb. Cortex 19, 1208–21. doi:10.1093/cercor/bhn164

Scullin, M.K., McDaniel, M.A., Dasse, M.N., Lee, J. hae, Kurinec, C.A., Tami, C., Krueger, M.L., 2018. Thought probes during prospective memory encoding: Evidence for perfunctory processes. PLoS One 13, 1–26. doi:10.1371/journal.pone.0198646

Scullin, M.K., McDaniel, M.A., Einstein, G.O., 2010a. Control of cost in prospective memory: evidence for spontaneous retrieval processes. J. Exp. Psychol. Learn. Mem. Cogn. 36, 190–203. doi:10.1037/a0017732

Scullin, M.K., McDaniel, M.A., Shelton, J.T., 2013. The Dynamic Multiprocess Framework: Evidence from prospective memory with contextual variability. Cogn. Psychol. 67, 55–71.

Scullin, M.K., McDaniel, M.A., Shelton, J.T., Lee, J.H., 2010b. Focal/Nonfocal Cue Effects in Prospective Memory : Monitoring Difficulty or Different Retrieval Processes. J. Exp. Psychol. Learn. Mem. Cogn. 36, 736–749. doi:10.1037/a0018971.l

Sellen, A.J., Louie, G., Harris, J.E., Wilkins, A.J., 1997. What Brings Intentions to Mind? An in Situ Study of Prospective Memory. Memory 5, 483–507. doi:10.1080/741941433

Shelton, J.T., Scullin, M.K., 2017. The Dynamic Interplay Between Bottom-Up and Top-Down Processes Supporting Prospective Remembering. Curr. Dir. Psychol. Sci. 26, 352–358. doi:10.1177/0963721417700504

Smith, R.E., 2003. The cost of remembering to remember in event-based prospective memory: Investigating the capacity demands of delayed intention performance. J. Exp. Psychol. Learn. Mem. Cogn. 29, 347–361. doi:10.1037/0278-7393.29.3.347

Smith, R.E., Bayen, U.J., 2004. A multinomial model of event-based prospective memory. J. Exp. Psychol. Learn. Mem. Cogn. 30, 756–77. doi:10.1037/0278-7393.30.4.756

Smith, R.E., McConnell Rogers, M.D., McVay, J.C., Lopez, J.A., Loft, S., 2014. Investigating how implementation intentions improve non-focal prospective memory tasks. Conscious. Cogn. 27, 213–230. doi:10.1016/j.concog.2014.05.003

Tulving, E., 1983. Elements of episodic memory. Oxford University Press, New York.

van der Wel, P., van Steenbergen, H., 2018. Pupil dilation as an index of effort in cognitive control tasks: A review. Psychon. Bull. Rev. 1–11. doi:10.3758/s13423-018-1432-y

Van Overschelde, J.P., Rawson, K. a., Dunlosky, J., 2004. Category norms: An updated and expanded version of the Battig and Montague (1969) norms. J. Mem. Lang. 50, 289–335. doi:10.1016/j.jml.2003.10.003

West, R., Carlson, L., Cohen, A.-L., 2007. Eye movements and prospective memory: what the eyes can tell us about prospective memory. Int. J. Psychophysiol. 64, 269–77. doi:10.1016/j.ijpsycho.2006.09.006

West, R., Scolaro, A.J., Bailey, K., 2011. When goals collide: the interaction between

prospective memory and task switching. Can. J. Exp. Psychol. 65, 38–47. doi:10.1037/a0022810

Zénon, A., Sidibé, M., Olivier, E., 2014. Pupil size variations correlate with physical effort perception. Front. Behav. Neurosci. 8. doi:10.3389/fnbeh.2014.00286