Running head: Prospective memory in the red zone

Accepted, Journal of Experimental Psychology: Applied, 20/02/2019.

Prospective memory in the red zone: Cognitive control and capacity sharing in a complex, multi-stimulus task

Luke Strickland[1,2], David Elliott[2], Michael David Wilson[1], Shayne Loft[1], Andrew Neal[3], & Andrew Heathcote[2]

[1] The School of Psychology,
The University of Western Australia, Australia

[2] The School of Medicine,
The University of Tasmania, Australia

[3] The School of Psychology,
The University of Queensland, Australia

Address for Correspondence

Luke Strickland,
School of Psychological Science, M304,
The University of Western Australia,
35 Stirling Highway,
6009 Perth, Australia
Email: luke.strickland@uwa.edu.au

Author Note

This research was funded by the Australian Government through the Australian Research Council (DP160101891). We thank Daniel White for programming the maritime surveillance task and for providing excellent technical support.

Abstract

Remembering to perform a planned action upon encountering a future event requires event-based Prospective Memory (PM). PM is required in many human factors settings in which operators must process a great deal of complex, uncertain information from an interface. We study event-based PM in such an environment. Our task, which previous research has found is very demanding (Palada et al., 2018), requires monitoring ships as they cross the ocean on a display. We applied the Prospective Memory Decision Control Model (Strickland, Loft, Remington & Heathcote, 2018) to understand the cognitive mechanisms that underlie PM performance in such a demanding environment. We found evidence of capacity sharing between monitoring for PM items and performing the ongoing surveillance task, whereas studies of PM in simpler paradigms have not (e.g., Strickland et al., 2018). We also found that participants applied proactive and reactive control (Braver, 2012) to adapt to the demanding task environment. Our findings illustrate the value of human factors simulations to study capacity sharing between competing task processes. They also illustrate the value of cognitive models to illuminate the processes underlying adaptive behavior in complex environments.

Keywords: prospective memory, linear ballistic accumulator model, unmanned aerial vehicle task, maritime surveillance

Public Significance Statement

We introduce a model that specifies the cognitive processes underlying Prospective Memory tasks, which require completion of deferred actions at some point in the future, and we apply it to a laboratory simulation of a complex, dynamic maritime surveillance task. Our model identifies multiple cognitive mechanisms that underlie performance in complex environments, with several clear implications for understanding and work design in practical settings.

Often, humans must remember to perform a planned action at some point in the future, a task referred to as Prospective Memory (PM; Einstein & McDaniel, 1990). In complex industrial-systems such as air traffic control and maritime surveillance with unmanned aerial vehicles (UAVs), PM tasks arise frequently, and are commonly reported as a source of human error (Dismukes, 2012). One common form of PM is event-based PM, in which the operator must perform a planned action when they encounter a particular environmental cue. For example, in maritime surveillance, some situations (e.g., military training operations) require the formation of temporary safety-zones under an operators' jurisdiction. In these circumstances, operators must remember to deviate from routine vessel classification, and instead remember that vessels matching particular features and travelling within the safety-zone must be reclassified and flagged for further evaluation (Nilsson, Van Laere, Ziemke, & Edlund, 2008).

Although field studies have been critical for identifying PM demands in complex workplace systems (e.g., Shorrock, 2005; Rothschild et al., 2005; Dismukes, Berman, & Loukopoulos, 2007), use-inspired laboratory experiments are also critical to understand the specific psychological mechanisms underlying the phenomena identified by field studies (Morrow, 2018; Stokes, 2011). For example, simulations of air traffic control have been used to identify factors that affect PM and to test the effectiveness of work design interventions such as memory aids (see Loft, 2014 for a review). However, applied PM research such as this has relied on verbal theorizing about the psychological mechanisms underlying PM. There are many potential advantages in extending this approach with quantitative "human performance modeling" (Byrne & Pew, 2009). For example, quantitative modeling can provide a unified account of disparate performance data, characterize the latent cognitive mechanisms underlying performance, and provide insights via simulation, for example by simulating decision making

within systems when human in-the-loop testing is not feasible. Recently, a quantitative model of event-based PM, “Prospective Memory Decision Control” (PMDC; Strickland, Loft, Remington, & Heathcote, 2018), was shown to provide a cohesive and informative account of the cognitive processes underpinning event-based PM in a basic laboratory paradigm.

PMDC may be a suitable basis for human performance modeling of event-based PM in a wide range of human factors applications. However, real-world cognitive demands can differ greatly from the demands of simple laboratory tasks. Notably, in previous studies, PMDC indicated that monitoring for PM items was possible at no cost to the information processing of concurrent tasks, suggesting that PM was able to rely upon surplus cognitive resources (Strickland et al., 2018). By contrast, in many settings of interest, such as aviation and defense, operators must integrate a great volume of complex and transient information from display interfaces and operate in dynamic and uncertain environments. Under such demands, cognitive resources may not remain in surplus. Indeed, practitioners are often particularly interested in identifying the level of cognitive demand that exceeds an operator’s capacity, because further demands beyond this point may impede the operator’s ability to perform within safety thresholds. This point is known as the 'red zone' (Wickens, Hollands, Banbury, & Parasuraman, 2015). To date, we do not know if or how PMDC can account for human behavior observed in the red zone. In the current study, we attempt to apply PMDC to performance in a simulated maritime surveillance task that is more representative of the type of complex task relevant to practitioners than the simple paradigm for which PMDC was originally developed. Critically, our paradigm is capable of raising task demand to the red zone (Palada, Neal, Tay, & Heathcote, 2018), and generalizing PMDC to complex tasks with high cognitive demand, thereby bridging the gap between basic cognitive research and practical applications of PM theory. Before

discussing our study further, we begin by introducing the general paradigm for studying event-based PM in the laboratory, and relevant previous findings.

Prospective Memory in the Laboratory

PM is typically examined in the laboratory using the Einstein and McDaniel (1990) paradigm. In this paradigm, participants perform an ongoing task (e.g., lexical decision, deciding whether strings of letters are words or non-words), and are asked to make an alternative task response at some point in the future (the PM task). In event-based PM, the PM task is to respond to items with certain target attributes (e.g., press an alternative key when you see a word containing the letter string containing 'tor'). Often, mean RTs to non-PM items are slower in PM blocks of trials than in control blocks, which is referred to as *PM cost* (e.g., Lourenço, White, & Maylor, 2013; Marsh, Hicks, Cook, Hansen, & Pallos, 2003; Smith, 2003). PM cost has been central to the development of PM theory, particularly to assessing potential trade-offs between monitoring for PM items and ongoing task performance (Smith, 2010; Einstein & McDaniel, 2010; Heathcote, Loft & Remington, 2015). The majority of detailed process modeling in the PM literature, which we review below, has focused on explaining the PM cost effect.

Verbal PM theories claim that PM cost indicates decreased ongoing task capacity due to capacity sharing between ongoing task processes and PM monitoring (Einstein et al., 2005; Smith, 2003). To test this claim, researchers have applied *evidence accumulation models* such as the "linear ballistic accumulator" (LBA; Brown & Heathcote, 2008) and the "diffusion decision model" (Ratcliff, 1978). These models provide a detailed process account of how humans make decisions, describing not only mean RT, but also the entire array of observed behavioral data for each participant. The models assume that evidence favoring each possible decision accrues towards threshold, and the first threshold to be breached determines the decision made. The

models estimate values of several latent psychological variables that underlie performance: *accumulation rates*, the speed at which evidence accrues; *thresholds*, the amount of evidence required to make decisions; and *non-decision time,* the time taken for other processes such as perceptual encoding of stimuli and motor responding.

To examine whether cognitive capacity for the ongoing task differs between PM block conditions and control block conditions, recent studies have fitted evidence accumulation models to ongoing task performance. Capacity sharing theories of PM cost predict that PM demands would have a detrimental effect on non-PM accumulation (Boywitt & Rummel, 2012; Horn, Bayen, & Smith, 2011). This follows from the notion in attention theories that processing speed is proportional to capacity (e.g., Bundesen, 1990; Gobell et al., 2004; Kahneman, 1973; Navon & Gopher, 1979; Wickens, 1980), as well as more recent findings that rates agree with other measures of cognitive capacity (Donkin, Little, & Houpt, 2014; Eidels, Donkin, Brown, & Heathcote, 2010), and that rates can be manipulated by increasing task demands (Logan, Van Zandt, Verbruggen, & Wagenmakers, 2014). In contrast, however, evidence accumulation modeling has consistently found there is no cost to non-PM accumulation rates under PM conditions (e.g., Ball & Aschenbrenner, 2018; Heathcote, et al., 2015; Horn & Bayen, 2015; Strickland, Heathcote, Remington, & Loft, 2017), a finding inconsistent with the capacity-sharing theories of PM. One recent study (Anderson, Rummel, & McDaniel, 2018) found some evidence of capacity sharing using the diffusion decision model, however, a better fitting LBA account of the data did not find such effects.

All accounts of PM cost to date using standard evidence accumulation models indicate that most of the PM cost effect results from an increase in ongoing task decision thresholds (Anderson et al.,2018; Ball & Aschenbrenner, 2018; Boywitt & Rummel, 2012; Heathcote et al.,

2015; Horn & Bayen, 2015; Horn et al., 2011; Horn, Bayen, & Smith, 2013; Strickland et al., 2017). Thresholds are the locus of strategic control in evidence accumulation models: individuals may raise thresholds globally to favor response accuracy over speed or raise the threshold of one choice relative to another to bias the decision process against that choice. Recently, Strickland, et al. (2018) incorporated this mechanism into PMDC, a process model of event-based PM. PMDC proposes that individuals increase ongoing task thresholds under PM conditions due to 'proactive control', which refers to anticipatory processes deployed in advance of a goal related event to assure an appropriate response to that event when it occurs (Braver, 2012). In the PMDC architecture, increasing ongoing task thresholds allows more time for PM accrual when processing PM items, increasing the likelihood of a PM decision (Heathcote et al. 2015; Loft & Remington, 2013). In addition to proactive control over thresholds, PMDC includes "reactive control", which refers to processes that occur just in time (i.e., when a PM trial is presented), during processing of a goal related event, to support appropriate responding to that event (Braver, 2012). Strickland et al. (2018) found that in the basic laboratory paradigms they examined, PM was supported by proactive and reactive control and not by capacity sharing between PM and ongoing task processing.

Prospective Memory in the Laboratory: Far from the Redline?

Although the reviewed studies failed to find capacity sharing between PM monitoring and ongoing task performance in the laboratory, there is reason to suspect capacity sharing may occur in other circumstances. Typical PM paradigms, including those applying evidence accumulation models, use simple ongoing tasks that may not fully occupy cognitive capacity, leaving 'reserve capacity' (Young, Brookhuis, Wickens, & Hancock, 2015) available to meet PM demands when necessary. This explanation is consistent with resource-based accounts of human

cognition, which specify that cognitive processes will only compete for resources when cognitive demands are high, and not when cognitive demands are low (e.g., Navon & Gopher, 1979; Norman & Bobrow, 1975). Indeed, participants report fewer task unrelated thoughts when they have PM demands, indicating greater on-task focus (Rummel, Smeekens, & Kane, 2017). In contrast, the types of complex, dynamic tasks that operators face in the real world can impose high workload, leaving little reserve capacity. Adding a PM load may then breach cognitive capacity limitations, leaving no option but to draw capacity from the ongoing task to support PM. In line with this, several simulated air traffic control studies have reported that PM load not only decreases the speed of aircraft acceptance and hand-off of aircraft and the speed of conflict detection (i.e., costs to ongoing air traffic management tasks), but can also impedes decision accuracy on ongoing tasks, for example, causing higher rates of missed conflicts (e.g., Loft, 2014; Loft, Finnerty, & Remington, 2011; Loft, Smith, & Remington, 2013; Loft & Remington, 2010; Loft, Smith, & Bhaskara, 2011).

The reviewed findings pose a challenge to the generalizability of basic PM paradigms to the real world. Evidence accumulation models offer by far the most detailed account of ongoing task performance data for simple laboratory tasks to date. However, in doing so, they indicate a lack of capacity sharing between PM and ongoing tasks in simple laboratory paradigms, and there is reason to suspect this lack of capacity sharing will not generalize to practical settings of interest. This calls for a careful analysis of when we would expect capacity sharing between PM and concurrent activities. As previously mentioned, the region in which additional task demands may consume capacity from other concurrent tasks has been referred to as the red zone (Wickens et al., 2015). The exact point at which additional task demands will lead not only to capacity sharing, but to rapid performance decrements and task failures is known as the ‘redline’ (Hart &

Wickens, 2010). We use the term ‘red zone’ to refer to the general region of task demand in which multiple tasks *may* compromise performance on single tasks, depending upon variability across individuals, and variability in task demand, whereas we use the term ‘redline’ to refer to the levels of task demand beyond which drastic task failures are very likely. System developers are often greatly interested in redlines, so that they can assure that their systems avoid them. However, redlines are difficult to identify in actual workplace settings, because humans will employ counter measures (e.g., get assistance from another operator) or adjust task processing strategies to avoid them (Loft, Sanderson, Neal, & Mooij, 2007). It is easier to identify redlines in the laboratory experiments where access to counter measures can be controlled. For example, Palada et al., (2018) recently identified and characterized the effects of a performance redline in a laboratory simulation of maritime surveillance. This paradigm forms the basis of the current study, and below we outline the relevant details of their paradigm, and how Palada et al. (2018) identified the red line level of demand.

Maritime Surveillance at the Redline

Palada et al. (2018)’s paradigm was designed to simulate the role of an unmanned aerial vehicle pilot monitoring the ocean with a camera view. Participants made decisions about multiple ships as they crossed the camera view (i.e., moved across the screen). They were required to classify ships as targets or non-targets based on the number of features on the ships. Ships with a certain number of features (e.g., more than four) were to be classified as targets, and ships with less features to be classified as non-targets. Several features of the task emulated realistic demands. First, the task emulated the naturalistic situation in which operators must respond to multiple stimuli presented together. Second, decisions were made in the presence of significant visual noise, emulating bandwidth limitations between ground control and

surveillance equipment. Third, participants were placed under time pressure, being required to make several responses before a 'trial deadline', the time at which all ships had crossed from one side of the screen to the other. Palada et al., manipulated two factors relevant to task demands: the number of ships participants needed to respond to per trial (trial load), and the amount of time participants had to respond to a group of ships (trial deadline). To measure the latent processes underlying their data, they fit the LBA, which provides a good account of decision making in this task despite it being applied to slower decision RTs than typically fit by evidence-accumulation models (see also Palada et al., 2016).

Palada et al. (2018) found that participants were able to maintain acceptable performance under high levels of trial load (5 ships on screen), with only a small increase in non-response rate (the number of ships not responded to before trial deadline). LBA parameter estimates indicated that participants did so by reducing the evidence required to make each decision (i.e., reducing threshold) as trial load increased. This control over thresholds implies strategic adaptations to task demand (i.e., cognitive control), rather than capacity-driven impairments from high task loads. However, participants were unable to employ strategies to adapt to tight trial deadlines. When deadlines were short – specifically, when there was only 6 seconds per trial – participants could no longer identify target ships reliably, with accuracy for target ships reducing to around chance. The LBA characterized these effects in terms of changes in accumulation rates. Under short trial deadlines, there were large increases in accumulation towards the incorrect decision for target ships. In contrast, non-target accuracy (correctly deciding that ships had less than the required number of features) remained in-tact at these deadlines. The costs to manifest accuracy, and latent accumulation, being specific to target ships may owe to the nature of evidence required for detection decisions. Detection decisions require the presence of a sufficient number

of target features, and lower capacity may lead to some features being missed. In summary, Palada et al. (2018)'s findings are consistent with severely reduced capacity for target detection, consistent with a capacity 'redline' for trial deadlines of 6 seconds.

The Current Study: Prospective Memory in the Red Zone

In the current study, we used the Palada et al. (2018) paradigm to study PM because it is a validated simulation of a complex and dynamic task that is also amenable to fine-grained analyses using evidence accumulation models. This paradigm is a micro-world simulation in which we can model performance under capacity demands that are potentially representative of a wide range of industrial settings. We do not instantiate PM beyond the 'red line' of maritime surveillance task demand, where demands would be so great that participants would fail at the primary task altogether. It is not clear what the practical or theoretical relevance would be of identifying capacity sharing between PM and ongoing tasks when primary task performance is already near chance. Instead, we add PM to an ongoing task to respond to 3 ships in 9 seconds, a level of demand that allowed adequate performance in Palada et al., (2018), but is still close to the 6 second 'redline' that they identified (which held whether participants were presented 2, 3, or 4 ships). In our paradigm, the three ships on each trial are presented concurrently. This multi-stimulus procedure may allow for potentially important phenomena that cannot be detected in single stimulus tasks (e.g., simultaneous encoding aspects of all three ships before orienting attention towards one specific ship). Our ongoing task uses a '7 feature rule', corresponding to Palada et al. (2018)'s most complex decision rule. This requires deciding whether ships travelling by the screen had 4 or more of the 7 possible features (target), or less than 4 (non-target).

Following the standard PM design, we include control blocks, in which participants only make ongoing task decisions on non-PM ships, and also PM blocks, that also include ships that require a PM response (PM ships). The PM task is to make an alternative response (e.g., right click one of two response boxes) instead of the ongoing task response (e.g., left click a response box), to ships which have both of two particular features (a life boat and a flag). We applied the PMDC model to understand the observed performance data. The reasons for this are twofold. First, PMDC can potentially provide a quantitatively adequate, and theoretically informative, account of performance in our task. Specifically, it enables examination of the latent psychological constructs underlying PM performance in the red zone. A more general aim is to examine whether PMDC can generalize to more complex and dynamic task environments. To this end, the study was designed to collect enough data (trials per participant) to provide robust fits using PMDC while maintaining a relatively low frequency of PM target presentations. We now introduce PMDC in detail and describe how it will apply to our experiment.

Testing Capacity and Control: Prospective Memory Decision Control (PMDC)

PMDC is an instance of Brown and Heathcote (2008)'s LBA, which assumes that evidence accumulates linearly for each possible decision until the total evidence of one accumulator reaches threshold, determining the response made. PMDC uses three racing LBA accumulators: two for the ongoing binary choice task responses (e.g., target and non-target) and a third for the PM task response. This architecture, as it applies to the maritime surveillance task, is depicted in Figure 1. The two-accumulator ongoing task variant of this model (applied to performance in control conditions) is similar to the LBA applied by Palada et al. (2018). The architecture models the decision-making process to each ship, with evidence accumulating towards three possibilities: that the ship is a target (has more than four features), a non-target

(less than four features), or is a PM ship (has both a life boat and a flag). For each decision, evidence begins at a random start point (uniformly in the range 0-*A*) and increases linearly as determined by a sample from a normal accumulation rate distribution with mean *v* and standard deviation *sv*. Accumulation to all possible responses runs independently until one of them reaches its threshold (*b*), and the first accumulator reaching threshold determines the response made. The total time to respond includes the time for the accumulation process plus a *non-decision time* constant. For each new ship for which a decision is made, evidence is reset and another decision process runs as just described. Below we describe in detail how, when applied to our task, the parameters of PMDC correspond to the degree of *capacity sharing* between PM and ongoing task processing, as well as the degree of cognitive control over PM and ongoing-task processes, in terms of Braver (2012)'s *proactive control* and *reactive control.*

Capacity sharing. We measure capacity sharing by comparing the accumulation rates to non-PM items (i.e., target and non-target ships without the PM feature configuration) in PM blocks with the accumulation rates to non-PM items in control blocks. PMDC estimates an accumulation rate towards the correct ongoing task decision (the 'match' accumulation rate, e.g., accumulation towards the target decision for target ships), and an accumulation rate towards the incorrect ongoing task decision (the 'mismatch' accumulation rate, e.g., accumulation towards the non-target decision for a target ship). Match and mismatch accumulation may be combined into two measures of overall processing (Palada et al., 2018), processing *quality* and processing *quantity*. Processing quality is given by match accumulation rate minus mismatch accumulation rate. Previous capacity sharing theories of PM cost have argued that capacity sharing would decrease the 'drift rate' parameter of the diffusion decision model, which indexes quality of processing (Boywitt & Rummel, 2012; Horn et al., 2011). Quantity of processing refers to the

sum of accumulation rates (match + mismatch). Note that decreased quantity of ongoing task processing under PM conditions could lead to slower ongoing task decisions, and thus also may potentially underlie the PM cost effect.

As reviewed, PM demands generally do not affect non-PM accumulation rates in simple paradigms, in terms of either the quality or quantity of ongoing task processing. However, in contrast to previous simple PM paradigms, the current study's demanding ongoing task may limit available cognitive capacity. Thus, in our study, PM conditions may induce a cost to accumulation to non-PM ships under PM conditions as compared with control conditions. This cost may affect processing quality, quantity or perhaps both.

Proactive control. PMDC proposes that participants may apply proactive control to increase ongoing task thresholds in PM blocks, so that when PM stimuli are presented ongoing task decisions do not pre-empt the PM decision. PMDC also includes control over the PM threshold, which may be adjusted based on factors such as the importance of the PM task, but our study does not manipulate such a factor. As reviewed, many studies find evidence for increased ongoing task thresholds in PM blocks, consistent with proactive control. Our task differs from previous studies in that it requires time-critical ongoing-task responding. Slow ongoing task responses can lead to failures to respond to all ships before trial deadlines, and Palada et al. (2018) found that with increased trial load participants reduced their thresholds in an attempt to meet trial deadlines. Therefore, if participants are concerned about non-responding, they may avoid proactive control over ongoing task decisions in our paradigm.

Reactive control. Reactive control occurs right when a PM related event is processed, to facilitate appropriate responding to that event. In the current context, this would be expected to occur when participants process PM ships (i.e., ships with both a life boat and flag). Under

PMDC, reactive processes do not affect thresholds, because thresholds are set prior to each decision beginning. Instead, reactive control affects accumulation rates. PMDC's reactive architecture is presented in Figure 2. As PM stimulus inputs are processed, they activate detectors for the possible decisions. For example, when PM items are processed, this can directly excite the PM accumulator (pathway A1 in Figure 2), leading to faster PM accumulation speed on PM items than non-PM items.

PM stimulus inputs can also inhibit the accumulation rates of ongoing task processes (pathways B1 and B2 in Figure 2), slowing ongoing task accumulation. This reactive inhibition would cause slower ongoing task accumulation when processing PM ships (which contain PM stimulus input), as compared with non-PM ships. Thus, to test for reactive inhibitory control, we compare ongoing task accumulation between PM ships and non-PM ships. Lower ongoing task accumulation to the PM ships indicates reactive inhibition. In simple paradigms, there is strong evidence for such reactive inhibition (Strickland et al., 2018), and it is critical to PM accuracy. We expect to replicate these reaction inhibition findings in the current study.

Multiple stimulus processes. Each trial presents participants three stimuli they must respond to before trial deadline (9 seconds). Although participants appear to make ongoing task decisions one at a time in our paradigm (Palada et al., 2016), there may be initial parallel encoding of the three ships. Moreover, participants will likely engage additional processes such as eye saccades to orient attention to the initial ship and orienting the motor response. To allow for this, we include a 'response order' factor in our analyses, which tracks whether each decision was the first, second, or third decision made within a trial. Following Palada et al. (2018), the model reported in text will include a different non-decision time for the first response relative to the other two. We expect to replicate Palada et al. (2018)'s finding that non-decision time is

longer for the first response than the others, and we examine whether this effect of response order can improve overall model fit.

Method

Participants

The study was approved by the Tasmania Social Sciences Human Research Ethics Committee. Thirty-six undergraduate students (12 females) from the University of Western Australia and the University of Tasmania participated in the study in exchange for course credit. Mean age was 21.6 (*SD* = 3.7). Due to a power failure, for one participant we did not record the data for one half of a control block (of 120 ship presentations). Another participant closed the program during a control block, resulting in missing data for the last 12 ships of that block. As we collected thousands of responses per participant, we were able to include the remaining data from these participants in all subsequent analyses. We excluded one participant from analyses entirely because they did not make a PM response to any of the 128 PM trials over the course of the experiment. One participant closed their practice trials early on both sessions, and another did not respond to a high proportion of practice trials (67% session one, 32% session two). Nonetheless, their data from the subsequent experimental trials appeared typical, so were retained for analysis.

Materials

Figure 3 depicts the maritime surveillance task environment. The task display comprised five horizontal channels of equal height. On each trial, three ships appeared and moved across the screen from right to left. The three ships were directly above and below each other (as in Figure 3), but the position of the group was otherwise random (i.e., the bottom ship could be in either the 1st, 2nd, or 3rd, lane). The ongoing task was to classify the ships based on how many

features each ship displayed. Figure 4 shows a ship with all seven possible features. Ships with four or more of these features were ongoing task targets, and ships with less than four of these features were ongoing task non-targets. Above each stimulus there were red and green response boxes, which participants left clicked to indicate their response (green for target, red for non-target). The PM task was to make an alternate response (right click) if a ship displayed both of two specific features (the flag and the life boat). The PM task always required right clicking the same response box, regardless of how whether the ship was an ongoing task target or ongoing task non-target (e.g., always right click the red response box for PM ships regardless of other ship features). Whether PM responses were right clicks to the red or green box was counterbalanced across participants (Table 1). We increased the difficulty of the task by obscuring each ship with a fog cloud overlay set at 50% opacity and by adding visual noise to the entire display.

To generate the stimuli, we created a list of all possible ships with 4 or more features (target ship list) and a list of all possible ships with less than four features (non-target ship list). Ship features were evenly represented across the ship lists (the ‘crane’ feature was just as likely as the ‘mast’ feature, and so forth). Both the target and non-target ship lists were further divided into two: PM ships (i.e., ships with both the flag and life boat features), and non-PM ships (i.e., ships that did not have both a flag and life boat). To fill the design as described below, ships were drawn randomly from these lists of ships, for example, for a control block, 126 non-PM target ships were drawn randomly from the non-PM, target ship list, and 126 non-PM non-targets were drawn from the non-PM, non-target ship list.

Design

Testing was split over two sessions. In each session, participants were given both PM and control 'blocks' of trials. Control blocks included 252 ships (84 trials), 126 non-PM target ships and 126 non-PM non-target ships. PM blocks included 252 ships, with 110 non-PM target ships, 110 non-PM non-target ships, 16 PM non-target ships and 16 PM target ships. Each block included a one-minute break in the middle, to reduce possible effects of fatigue. Stimuli numbers were split evenly across each half of the block for each stimulus type. Within each half block, stimulus presentation order was randomized. To collect a sufficient number of PM observations, each of the two sessions of the experiment included two PM blocks and one control block. Across the two PM blocks we observed responses to 504 ships (168 trials), 64 of which were PM ships (32 PM targets, 32 PM non-targets).

The two PM blocks were always presented sequentially. Thus, the order in which control and PM blocks were presented for each session was either control-PM-PM or PM-PM-control. Assignment of block order to session was counterbalanced across participants, resulting in two possible block order schemes: either control-PM-PM/PM-PM-control or PM-PM-control/control-PM-PM. In summary, participants were assigned to one of four conditions (see Table 1) which determined: (1) the order in which they performed experimental blocks (PM-PM-control vs. control-PM-PM) and (2) the response button used for PM responses (red vs. green). These factors were orthogonally counterbalanced as displayed in Table 1. Across 32 participants, we achieved a full counterbalance: an equal distribution of subjects in each cell of our counter-balancing scheme (with eight participants in each condition as can be seen in Table 1). Across our two testing sites we ended up with 3 more participants than planned for, resulting in an additional 2 participants in counterbalance group 1, and 1 extra participant in counterbalance

group 2. Our analyses below included all our data, but the conclusions held even if we excluded the last three participants tested (for a full counterbalance).

Procedure

Training required participants to read through a slideshow of task instructions. They were informed that they were required to classify ships as either targets or non-targets using a 7-feature classification rule. Under this rule, ships with 4 or more features needed to be classified as target by left-clicking the green response box above that ship. Ships with less than 4 features were to be classified as non-target by left-clicking the red response button above that ship. They were told that multiple ships would appear on-screen in adjacent positions, and that the ships would enter and leave the screen at the same time. Participants then practiced classifying ships. For each response, they received feedback: a red cross next to the ship for an incorrect response, or a green tick next to the ship for a correct response (no feedback was provided in the experimental phase). Participants completed 36 training trials (108 ships) each testing day.

After training, participants received instructions for the experimental blocks. Participants were informed that the experimental phase would last 40 minutes. For control blocks, participants were instructed to perform the target/non-target classification task. For the PM blocks, participants were instructed to continue performing the target/non-target classification task; however, they were also instructed that if a ship had both a life boat and a flag, they should right click on the green/red response box (which depended on the counterbalance of the participant) above the ship instead of making a left click classification response. We chose the life boat and flag features because they were both quite distinct from the grey noise overlay added to the task. This avoids a situation where the PM task required greater perceptual effort than the ongoing task. After receiving PM instructions, participants then viewed an image of a

ship with labels and arrows indicating the life boat and flag features (similar to Figure 4). As has been common practice in the PM literature, we included a 'filler' task after the PM instructions, before the commencement of the ongoing task. Before beginning the experimental blocks, participants were told that they would have 2 minutes to complete a puzzle using pen and paper. After the puzzle, they began the experimental block.

Results

Before discussing the PMDC model fits to our data, we report conventional analyses of RTs and accuracy. Our analyses were conducted with the use of the R programming language (R Development Core Team, 2018). We used linear mixed models, the recommended method for repeated measures data (Pinheiro & Bates, 2000) as implemented in the R *LME4* package (Bates, Mächler, Bolker, & Walker, 2015). Post hoc analyses were performed using *glht* from the R MULTCOMP package (Hothorn, Bretz, & Westfall, 2008) with Tukey adjusted p-values. We analyzed PM condition (PM block, control block), stimulus type (non-PM target ships, non-PM non-target ships, PM target ships, PM non-target ships), response order (response 1, response 2, response 3) and session (session one, session two) as fixed effects, and subject as a random effect. All responses with RT < 0.2 s (0.03%) were excluded from subsequent manifest and model analysis.

Separate models were fit for ongoing task RTs, ongoing task accuracy, PM task RTs, PM task accuracies, and PM false alarms. Each model was built by including each factor separately and comparing against a null model (with subject as a random effect) to check for significant (p <.05) improvement in model fit (using Wald chi-square tests). We built each model by stepwise addition of each significant factor, retaining those which provided significant improvement in model fit as a main effect or interacting with other factors. Detailed results of the Wald chi-

square tests and post hoc comparisons are provided in the supplementary materials. AIC-based model selection yielded the same models to the method described. In text, we focus on effects that were statistically significant, unless specified otherwise.

Ongoing task responses were scored as correct if the participant correctly identified the stimulus. PM responses were scored as correct if the participant right-clicked the stimulus on a PM trial. Occasionally (0.31% of responses to non-PM ships, 11.9% of responses to PM ships), participants right-clicked a stimulus, but on the response box which they were not assigned for their PM task (e.g., a participant told to right click green for PM right clicking red instead). These were scored as PM responses, on the assumption that this was more likely to reflect a procedural response error (in which participants intended to respond PM), than a PM error.

False Alarms

PM false alarms, that is PM responses to non-PM ships, occurred to 1.4% of non-PM ships in PM blocks. The false alarm rate decreased from session one (M = 2.1%) to session two (M = 0.6%). We observed some false alarms in control blocks (14 responses, 0.08% of the total observed), despite participants being instructed before control blocks that they did not have to monitor for PM items. More than half of these (8) came from a single participant. No other participant made more than one false alarm during control blocks. As there were very few of these false alarm responses in control blocks, they are excluded from all further analysis, including the subsequent PMDC model fitting.

PM Costs

The results of PM cost analysis are presented in Table 2. For the ongoing task (feature classification), correct RTs for target and non-target ships in PM blocks were longer than in control blocks, except for non-target ships in the first session. Accuracy was consistently lower

in PM blocks for both target and non-targets; however, the reduction in accuracy for targets in the first session (7%) was substantially greater than non-targets in session one (1.3%), targets in session two (0.8%), and non-targets in session two (0.7%). Responses to target ships were less accurate than responses to non-targets, except for the control block in the first session, where there was no significant difference in accuracy between target and non-target ships.

PM Task Performance

Other than response order (which will be covered in the next section) only "session" affected correct PM RTs, which decreased from the first session (M = 1.90, SE = .03) to the second session (M = 1.72, SE = .03). PM task accuracy for targets and non-targets is shown in Table 3. PM accuracy was higher for target ships than for non-target ships both sessions. PM accuracy to target ships decreased from session one to session two.

Response Order

To determine whether response order within a trial affected performance, we compared correct RTs and accuracy of the first, second and third responses. RTs and accuracies by response order are presented in Table 3. For the ongoing task, the first RT on each trial was longer in both control and PM conditions, but there was no significant difference in RTs between the second and third response in either session. For the PM task, correct RTs followed a similar pattern to the ongoing task: the first response took substantially longer, but there was no difference in RT between the second and third response. For the ongoing task, accuracy decreased with response order, except for PM blocks in session two, where accuracy for the first

and second ongoing task response did not significantly differ. Overall, response order did not affect PM accuracy.

Consistent with the higher mean RT on the first response of each trial, we suspected that individuals were engaging in additional non-decisional processes at the start of each trial (Palada et al., 2018), as a result of parallel encoding of stimuli, motor response orientation and eye saccades. In order to justify the inclusion of a separate non-decision time parameter for the first response in the subsequent PMDC modeling, we first checked for further differences in the manifest data, by examining the 0.1 quantiles of RT for each participant. Because non-decision time is a constant time added to each trial, the fastest responses (i.e., those with relatively small decision times), should be sensitive to shifts in non-decision time. We found that the 0.1 quantiles of RTs were longer for the first responses on each trial (M = 1.17, SE = .016) relative to other responses, suggesting that we should include a difference in non-decision time in subsequent model analysis. In contrast, there was only a small difference between the second (M = 0.95, SE = .003) and third (M = 0.94, SE = .003) responses.

To conclude, we found PM costs not only to ongoing task RT, but also to ongoing task decision accuracy. This is consistent with findings from PM in simulations of air traffic control (e.g., for a review see Loft, 2014). These data suggest that our simulation of maritime surveillance is sufficiently cognitively demanding to place participants in the ‘red zone’, in which PM monitoring and ongoing task performance must trade-off. In contrast to previous studies, our design is sufficiently high powered and controlled to also apply the PMDC model, which allows us examine the cognitive control and capacity related mechanisms that underlie task processing trade-offs and PM errors in the red zone.

Model Results

We apply the PMDC model, as described in the introduction and depicted in Figure 1. We report thresholds in terms of $B > 0$ (equal to b - A). Our design included several factors that parameters could conceivably vary over: PM condition (control vs PM), stimulus type, latent response accumulator (non-target, target, PM), response order, and session of experiment. In order to constrain the flexibility of the model so that it would remain a 'measurement model' with parameters that can be accurately estimated from the data, we applied several *a priori* constraints on which parameters were allowed vary with design factors.

We constrained A to be the same across all accumulators and all conditions. To account for the possibility that participants took extra time to encode the locations of the three ships in parallel, and to orient their attention and motor response (mouse cursor) to the first ship, we allowed a different non-decision time for the first response in each trial relative to the other response orders. Consistent with Palada et al. (2018), we did not allow any additional flexibility over response order, as doing so would risk too little data for acceptable parameter estimation (recall that we only observe responses to 64 PM target ships and 64 PM non-target ships for each participant). We fixed non-decision time over all other factors. We estimated thresholds separately for each accumulator, for each PM condition (PM vs. control), and for each session of the experiment. Following results and precedent from simple laboratory paradigms (Strickland et al., 2017, 2018), we relied on thresholds alone to capture the effects of session. We did not estimate separate thresholds in response to stimulus type. As stimuli were randomly presented, estimating thresholds based on stimulus type would be circular: if participants could change adjust thresholds in contingent on stimulus identity, then they must have already known what that stimulus would be, and the decision process would not need to occur. We allowed mean

accumulation rates to vary over PM condition and stimulus type. As is common in applications of the LBA, we included only two accumulation rate standard deviations: a 'matching' standard deviation parameter for correct decisions and 'mismatching' standard deviation parameter for errors. We fixed the standard deviation of the 'mismatch' accumulator at 1, so that the model was identifiable (Donkin, Brown, & Heathcote, 2009). From these aforementioned constraints, the 'top' most flexible model we fit comprised 29 free parameters – one start-point noise parameter, 10 thresholds, 15 mean accumulation rates, 1 standard deviation for 'match' accumulation rates, and 2 non-decision times.

Sampling

We estimated model parameters using Bayesian estimation, which produces a probability distribution for each parameter value. We used the Differential Evolution Markov Chain Monte Carlo algorithm (DE-MCMC; Turner, Sederberg, Brown, & Steyvers, 2013), an algorithm that is effective in Bayesian estimation even when parameters are highly correlated, as implemented in the DMC suite of R functions (Heathcote et al., in press). Bayesian estimation requires specifying 'prior' distributions, which specify beliefs about parameter values prior to modelling the data. As this was the first application of PMDC to our task, we specified fairly non-informative priors, which are listed in Table 1. Our priors were similar to the priors used by Strickland et al., (2018). However, in order to account for the longer RTs observed in our task, we doubled the prior means of the *B* parameters (the amount of evidence required for a decision) and increased the upper bound on non-decision time to 3 seconds.

The DE-MCMC algorithm requires running many parallel chains, which share information to efficiently converge to the posterior distribution. We ran three times as many chains as parameters (e.g., for the most complex model with 29 parameters we ran 87 chains). To

reduce computational requirements, we 'thinned' during sampling, only retaining one out of every twenty iterations that we ran. We ran 3600 iterations at a time, retaining 180 of them after thinning. We continued to run more iterations (replacing the samples from the previous 180) until all samples were stationary, mixed, and converged, which we evaluated both with Gelman's $\hat{R}$ statistic (Gelman et al., 2014) and with visual examination of trace plots of the samples.

Model Fit

Before interpreting the model, we examine whether it adequately accounts for the observed data, using graphs of 'posterior predictive' (Meng, 1994) model fit. This involves using the posterior parameter samples to simulate data from the model and comparing them with the actual data. Figure 5 compares the fit between simulated and observed data across the different PM conditions, stimulus types, and responses. Overall, the PMDC architecture provided a good fit to accuracies and RTs of the ongoing task and of the PM task. We separately examined fit to the 'response order' factor, depicted in Figure 6. The model adequately captures the major effect of 'response order' on RT. There is some miss-fit to the extended tail of RTs for response order one, and slight miss-fit for some of the accuracy trends. Although this miss-fit may be addressed with more parameters[1], it is not very substantial, and we decided that further gains in fit to our data would not be worth the added model complexity (i.e., more potential to over-fit to noise in our data and thus not generalize well to future data). Next, we describe how the model achieved good fits to our observed data in terms of latent psychological quantities. To do so, we first explore the possibility of constraining the model further with model selection.

[1] We experimented with more complex models (e.g., thresholds and rates varying by trial position, 78 parameters) and found that they could fit response order effects more closely, for example helping in fitting the longer 0.9 quantile for first responses. Although these models added extra mechanisms by response order, their average effects across PM conditions were very similar to those reported in text.

Model Selection

In addition to the *a priori* constraints on the model, we tested whether the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) suggested any further constraint would be appropriate. DIC measures how 'good' a model is by considering both complexity and goodness of fit to the data. A lower DIC indicates a better model. DIC values were summed across participants. We tested three models. The first was the 'top', most flexible model, that allowed both capacity sharing and proactive control in the PM and control conditions. The top model had 29 parameters and a DIC value of 110468. We tested an 'only capacity sharing model', in which thresholds were fixed across PM conditions, precluding proactive control. Although this model was simpler, with only 25 parameters, it had a substantially larger DIC than the top model (DIC difference = 1140), indicating support for the top model. We tested an 'only proactive control' model, in which accumulation rates to non-PM stimuli did not vary across PM and control conditions. This model had 25 parameters and was also rejected in favor of the top model, (DIC difference = 322), suggesting that changes in non-PM accumulation rates were required to fit the data. Thus, overall, the model selection is in favor of both thresholds and rates varying across PM and control blocks. We now test the direction and magnitudes of these variations in thresholds and rates.

Model Summary

Parameters

To summarize model parameters, we created a group-averaged posterior distribution, which averaged every posterior sample across participants. To test differences between parameters, we took the differences between the parameters for each posterior sample, and then averaged the result across participants, giving a group averaged posterior difference distribution.

To summarize the difference distributions, we report the proportion of posterior samples of the difference that were above or below 0 (*p*). So that *p* values are consistent (with smaller *p*s evidencing an effect), we report the posterior *p* against effects running in the observed direction. Thus, if a tested difference in parameters is most often sampled above 0, we report the *p* that it is below 0, and if a difference is most often sampled below 0, we report the *p* that it is above 0. We also report *Z*, the mean of the difference distribution divided by the standard deviation (similar to a *Z* score because the difference distributions of posterior samples are approximately normal). We include the latter because many of our observed effects were 'significant' in that $p = 0$, and so we needed some way to compare their relative size. For *t0*, we found an effect of position in trial, with the *t0* to the first response on each trial much higher ($M = 0.77$s, $SD = 0.004$) than *t0* to the subsequent two ($M = 0.12$s, $SD = 0.003$s), $Z = 222.7$, $p = 0$. This non-decision time effect is consistent with greater encoding time at the start of each trial, in which participants have to locate the three ships, and perhaps have to move their mouse to the first ship. The values of non-decision time are similar to those reported by Palada et al., (2018). The posterior mean of the standard deviation in rates for the matching accumulator was 0.65 ($SD = 0.01$). As the standard deviation for the mismatching accumulator was fixed at 1, this replicates most previous modeling with the LBA where the match accumulator is generally found to have lower variability than the mismatching accumulator. The posterior mean of the *A* parameter was 2.71 ($SD = 0.05$).

Capacity Sharing

We tested capacity sharing by comparing accumulation rates to non-PM ships across control blocks and PM blocks, with slower non-PM accumulation in PM blocks indicating capacity sharing. Figure 7 shows the accumulation rates for non-PM ships broken down by accumulator type (matching vs. mismatching) and stimulus type (target vs. non-target). The rate

for the matching accumulator was lower for targets in PM blocks (M = 2.32, SD = 0.03) than in control blocks (M = 2.48, SD = 0.03), Z = 7.37, p = 0 (top right panel). Although the effect was smaller, the rate for the matching accumulator was also lower for non-targets in PM blocks (M = 2.25, SD =0.03) than control blocks (M = 2.30, SD =0.03), Z = 2.45, p = .007 (top left panel). The rate for the mismatching accumulator was not appreciably different in PM blocks *(M* = -0.30, SD = 0.05) than control blocks (M= -0.33, SD = 0.06), Z = 0.36, p = .361, when targets were presented (bottom right panel). However, the rate for the mismatching accumulator was marginally higher for non-targets in PM blocks (M = -1.07, SD = 0.07) than control blocks (M = -1.22, SD = 0.08), Z =1.6 p = .06 (bottom left panel).

We also combined the rate measures to test changes across condition in processing quality (match accumulation – mismatch accumulation) and processing quantity (match accumulation + mismatch accumulation) for non-PM ships. We found lower processing quality under PM conditions for both targets, Z = 2.66, p = .004, and non-targets, Z = 2.06, p = .02. We found lower processing quantity for targets, Z = 1.98, p = .03, but not non-targets, Z = -1.08, p = .14. Overall then, with three effects on accumulation rates consistent with lower ongoing task capacity, and one rate not affected by PM condition, these results suggest lower ongoing capacity in PM blocks than control blocks. These effects were modest in size compared to some other effects in the model, but in the ‘Model Exploration’ section we will demonstrate that they were important in accounting for PM costs.

Proactive Control

To test for proactive control, we compared ongoing task thresholds in PM blocks with control blocks; with higher ongoing task thresholds in PM blocks being indicative of proactive control. Thresholds are plotted in Figure 8. Thresholds were higher to make target decisions in

PM blocks (session one M = 2.32, SD = 0.03, session two M = 2.03, SD = 0.03) than control blocks (session one M= 2.00, SD = 0.03, session two M = 1.91, SD = 0.03) for both session one, Z = 10.92, p = 0, and session two, Z = 4.38 p = 0. Proactive control over non-target decisions was less apparent. In session one, thresholds to make non-target decisions were not higher in PM blocks (M = 2.21, SD =0.03) than control blocks (M = 2.21, SD = 0.04), Z = 0.16, p = .434. However, in session two, non-target thresholds were higher in PM blocks (M = 2.04, SD = 0.03) than control blocks (M = 1.91, SD = 0.03), Z = 4.69 p = 0. Thresholds towards the PM decision were not substantially higher for session one (M = 1.73, SD = 0.04) than for session two (M = 1.73, SD = 0.04), Z = 0.28, p = .39.

Reactive Control

Excitation. In order to examine reactive processes, we compare differences in accumulation between PM ships and non-PM ships. Trivially, as would be expected with reasonable PM accuracy, PM accumulation was substantially faster to PM ships (target ships M= 2.14, SD = 0.04, and non-target ships M =1.70, SD = 0.04) than to non-PM ships (M = -3.79, SD = 0.12), Z = 52.64, p =0. More interestingly, PM excitation was greater to ‘target’ PM ships (PM ships that also satisfy the ongoing task target detection rule), than to ‘non-target’ PM ships (PM ships that do not satisfy the ongoing target detection rule), Z = 11.5, p = 0.

Inhibition. We tested for reactive inhibition by comparing ongoing task accumulation rates between PM ships and non-PM ships. These rates are plotted in Figure 9. Lower ongoing task accumulation to PM ships indicates inhibition. For target ships, accumulation towards the decision to respond 'target' was much lower when the target was a PM ship (M = 0.61, SD =0.08) than when it was a non-PM ship (M = 2.32, SD =0.03), Z = 24.79, p = 0 (top right panel, Figure 9). Similarly, for non-target ships, the speed of accumulation towards the non-target decision was

much lower when the non-target was a PM ship (M = 0.88, SD = 0.05) than when it was a non-PM ship (M = 2.25, SD = 0.03), Z = 29.64, p = 0 (top left panel). For target ships, there also was evidence that the error accumulation rates were lower when the target was a PM ship (M = -1.56, SD = 0.15), than when it was a non-PM ship (M = -0.30, SD = 0.05), Z = 8.18, p = 0 (bottom right panel). For non-target ships, error accumulation was also lower when the non-target was a PM ship (M = -1.44, SD = 0.14) than when it was a non-PM ship (M =-1.07, SD = 0.07), Z = 2.53, p = .004 (bottom left panel).

We also examined inhibition in terms of processing quality and processing quantity. For target ships, ongoing task processing quality was lower for PM ships than non-PM ships, Z = 2.64, p = .005. For non-target ships, ongoing task processing quality was lower for PM ships than non-PM ships, Z = 6.5, p = 0. For target ships, ongoing task processing quantity was much lower for PM ships than non-PM ships, Z = 17.58, p = 0. This was also the case for non-target ships, Z = 11.36, p = 0.

Model Exploration

In complex cognitive models, the contribution of model processes to the observed data can be difficult to discern. Here, we seek to understand the model processes that underlie PM cost and accuracy by removing them from the model and examining the miss-fit to PM cost and PM accuracy that follows. To the extent that removing a model mechanism causes miss-fit to an effect, that effect can be ascribed to the mechanism in the full model. We summarize model fit by examining the percentages of PM accuracy and PM cost predicted. This percentage can be over 100 if the model actually predicts a larger number than the data.

PM accuracy. First, we tested the contributions of proactive and reactive control to PM accuracy. Figure 10 depicts the results. We removed all control over ongoing task decisions from

the model by setting thresholds in PM blocks to control block levels (removing proactive control of ongoing tasks decisions), and also setting ongoing task accumulation rates on PM ships to the level on non-PM ships (removing reactive control). Whereas the full model predicted 98% of PM accuracy to target ships and 103% of accuracy to non-target ships, the model with control removed predicted only 59% of accuracy to target ships, and 65% to non-target ships. Next, we examined a model with proactive control in the model, but not reactive control. This predicted 65% of PM accuracy to target ships, and 66% of accuracy to non-target ships indicating that proactive control made some contribution to PM accuracy in the full model. However, allowing reactive control into the model, but not proactive control, resulted in much better fits to PM accuracy, with 95% of PM accuracy predicted to target ships and 100% of accuracy predicted to non-target ships. This illustrates that for the current task, proactive control had relatively little influence on PM accuracy, whereas reactive inhibitory control was critical.

PM cost. Next, we tested the extent to which the PM cost effect could be accounted for by proactive control compared to capacity sharing. We examined a model with proactive control removed by setting all ongoing task thresholds to the values in control conditions; and we examined a model with capacity sharing removed by setting all non-PM accumulation rates to the values from control blocks. The results are displayed in Figure 11. We only discuss cost to responses to target ships, because cost to non-target ships was small, and the 95% credible intervals of all the models overlapped with the true effect. The full model predicted 104% of cost to target ships, whereas removing capacity sharing resulted in predicting only 54% of cost to target ships. Similarly, removing proactive control predicted only 39% of cost to target ships. These results suggest that both capacity sharing and proactive control played substantial roles in the observed PM cost effects to target ships.

Individual Differences in PM Accuracy

We were also interested in how the model explained variation in PM accuracy across participants. We investigate this by correlating model mechanisms with PM accuracy. To obtain correlations, we calculated model mechanisms (e.g., proactive control over target decisions) for each posterior sample for each participant, and then calculated the correlation for that posterior sample. The result is a distribution of posterior correlations (Ly, Boehm, et al., 2018). We applied a correction to the correlations, so that they are suitable for inference about the population rather than the sample (Ly, Marsman, & Wagenmakers, 2018). We plot the posterior means, and credible intervals, of the resulting correlation distributions in Figure 12. Below we report in text the posterior means of the substantial correlations.

In our previous work using a simple paradigm, we found that only reactive inhibition towards the 'match' ongoing task decision strongly correlated with PM accuracy. We replicated this reactive inhibition effect here, for PM accuracy both to target PM ships (r = .74) and to non-target PM ships (r = .73). Furthermore, reactive inhibition to the non-target response for non-targets was correlated with target PM accuracy (r = .55), and reactive inhibition of the target response to target PM ships was correlated with non-target PM accuracy (r = .57). These correlations suggest that reactive control ability is a strong determinant of individual differences in PM accuracy.

We found another parameter that correlated with PM accuracy here that did not correlate in the simple PM paradigm: the PM accumulation rate. The speed of PM accumulation correlated with PM accuracy both to target PM ships (r = .55) and non-target PM ships (r = .82). Furthermore, the PM accumulation rate towards non-target PM ships was correlated with PM

accuracy to target PM ships (r = .62) and the PM accumulation rate towards target PM ships was correlated with PM accuracy towards non-targets (r = .49).

Multiple Stimulus Encoding

In each trial, participants were presented three ships and were instructed to respond to in any order they wished. As reviewed above, we found substantially slower responses for the first ship on each trial, consistent with Palada et al. (2018). This motivated a subsequent analysis of our manifest data. Although it is unlikely that participants would switch their attention between the ships during decisions (Palada et al., 2016), it is possible that at the beginning of each trial, prior to focusing attention on the first ship (and performing evidence accumulation), participants might either perform some initial scan of the ships for PM features, or notice the PM features during initial encoding of the locations of the three ships. We tested this for each participant by performing a chi-square test on the frequency with which PM items were responded to first. As PM ships were interspersed randomly amongst the other ships, the null hypothesis was that PM ships would be responded to first no more than chance. The null was calculated for each participant by examining the total number of responses recorded first, second, and third within a trial (this was not exactly a third due to non-responses, although very close). We tabulate the results of the chi-square tests for each participant in the supplementary materials.

We found four participants who responded to PM ships first more often than chance (p less than the family-wise error rate of .0014). This motivated two checks of our model results. First, we checked whether the inferences reported above held without the four ‘PM first’ participants included in the data, and we found that they did. Second, we examined non-decision times of the four ‘PM first’ participants compared with the rest of the participants. We found that the non-decision time for the first response on each trial was slower for these four participants

(M = 0.84s, SD = 0.009s) than other participants (M = 0.76s, SD = 0.004s), Z = 7.46, p = 0, whereas the non-decision times for the other two responses were slightly faster for the four participants (M = 0.11s, SD = 0.004s) than for the rest of the sample (M = 0.12s, SD = 0.003s), Z = 2.99, p = .006. Given these results, it is possible that participants who encode stimuli more thoroughly at the beginning of the trial can notice PM ships during encoding, and then subsequently orient their attention to a PM ship first for that trial.

Discussion

In the current study, we examined PM in a laboratory simulation of a complex, dynamic, maritime surveillance task. We found substantial PM error rates in our task, similar to those observed in simple laboratory paradigms (Einstein & McDaniel, 1990). Replicating findings from simple laboratory PM tasks, we found a PM cost to RT of the ongoing target detection task. Specifically, in PM blocks of trials we found slower RTs to (non-PM) ongoing task targets. We also found a PM cost to ongoing-task target accuracy. We used our data to test the recently developed PMDC model (Strickland et al., 2018). We found that PMDC fitted well to the observed performance data, despite the RTs in this task being much slower than in tasks PMDC has previously been applied to. This finding corroborates other recent studies that have successfully validated accumulation models in tasks that require longer decisions, and thus are more representative of decisions in complex real-world task environments (Lerche & Voss, 2017; Palada et al., 2016, 2018). Critically, we found that PMDC was applicable to, and informative about, human performance in a ‘red zone’ of task demand. PMDC indicated capacity sharing between PM monitoring and ongoing task performance, contrasting with previous findings from simple paradigms that PM monitoring does not impact ongoing task capacity (Strickland et al., 2018). We also found support for proactive and reactive control processes

underlying PM, generalizing previous findings (Strickland et al., 2018) to the current more complex paradigm.

Capacity sharing. We found evidence for capacity sharing between PM monitoring and performing the ongoing task. Accumulation rates to non-PM ships in PM blocks suffered as compared with accumulation in control blocks. For non-PM ships that were ongoing task targets, there was a cost for both processing quality (lower quality ongoing task accumulation under PM conditions) and quantity (less quantity of ongoing task processing under PM conditions). For non-PM, non-target ships, we found only a cost to processing quality, and not quantity. This asymmetry in PM cost across ongoing task targets and non-targets is also evident in the manifest data. PM cost to RT was more substantial for target ships than non-targets, and PM cost to ongoing task accuracy was only present for target ships. The asymmetry also appeared in Palada et al. (2018)'s finding that the point of capacity 'redline' reduced target ship accuracy to chance but left non-target ship accuracy intact. It appears that in this maritime surveillance paradigm, capacity costs primarily affect ongoing task target detection, rather than target rejection.

At first glance, our finding of capacity sharing in maritime surveillance may seem consistent with early verbal PM theories, which assumed the PM cost effect in more simple PM tasks is driven by capacity sharing (Einstein et al., 2005; Smith, 2003), and starkly contrasting with more recent research applying evidence accumulation models that have consistently failed to find evidence of capacity sharing in simple PM tasks (e.g., Ball & Aschenbrenner, 2017; Heathcote et al. 2015; Horn & Bayen, 2015; Strickland et al., 2017, 2018). However, we do not believe our findings indicate that capacity sharing undergirds PM in all situations. Instead, our findings suggest that capacity sharing is likely driven by the demands of the task environment. According to resource-based theories of human cognition (e.g., Navon & Gopher, 1979; Norman

& Bobrow, 1975), we can only identify whether two processes share capacity in situations where capacity cannot be held in reserve. As outlined in the introduction, our maritime surveillance task is substantially more demanding than the simple tasks typically used in PM research. Previous work by Palada et al. (2018) places our level of task demands in or close to the 'red zone' (Wickens et al., 2015), where participants are likely to have insufficient 'reserve capacity' to meet additional task demands, resulting in capacity sharing between tasks. Thus, the capacity sharing finding we observed in the present task likely was driven by the demands of the task. Moreover, this demonstrates that effects which appear similar across simple and complex tasks using coarse measures (e.g., mean RT), can be driven by very different psychological processes, with different implications for psychological theory and practice.

Additionally, owing to the fact that PM accumulation rate can also be taken as a measure of PM capacity, the results of the present study indicate that individual differences in capacity are associated with PM accuracy. Previous work applying PMDC to simple paradigms found that PM accuracy only correlated across subjects with inhibition of the ongoing task (Strickland et al., 2018). However, in the present study we found that the PM accumulation rate also correlated with PM accuracy across participants. It is likely that Strickland et al. (2018) did not detect these individual differences because participants had sufficient capacity to perform both the ongoing task and maintain PM monitoring without capacity sharing, resulting in PM accuracy being largely determined by cognitive control over ongoing task processes. Thus, in the current paradigm where spare capacity is limited, it follows that individual differences in PM capacity could explain differences in PM accuracy.

Proactive control. Replicating work from simple paradigms (e.g., Ball & Aschenbrenner, 2018; Heathcote et al., 2015; Horn & Bayen, 2015; Strickland et al., 2017,

2018), we found that participants increased their ongoing task thresholds in PM conditions as compared with control conditions, consistent with 'proactive control' over ongoing task decisions. The strongest effect was on thresholds to make 'target' decisions. These threshold increases accounted for a substantial portion of PM cost. However, our model exploration revealed that although these threshold increases were important to explaining PM cost, they only had minor benefits to PM accuracy. There was also some proactive control over thresholds to make 'non-target' decisions, however this effect was smaller, only appeared on session two of the experiment, and did not produce any substantial PM cost to non-targets, or any substantial benefits to PM accuracy.

All in all, ongoing task threshold control appears less relevant to PM accuracy in the current study than it was in our previous application of PMDC (Strickland et al., 2018). Proactive control may be more limited in the current paradigm due to the trial response deadlines. Proactive control would increase RTs and risk breaching deadlines, and Palada et al. (2018) reported that participants decreased their thresholds to avoid non-responses. Thus, it is possible that without response deadlines, participants could exert more proactive control, and this control would account for more variance in PM accuracy, perhaps also discouraging the need for capacity sharing. This limitation on proactive control is important to note for real world applications of PMDC. Often, operators face time pressure to respond, either in time to avoid an adverse event, or because they need to switch to other tasks. Such situations may render a proactive control strategy ineffective, potentially compromising PM accuracy, or increasing capacity demands of PM.

Reactive control. Unsurprisingly, we found larger PM accumulation rates for PM ships than non-PM ships, consistent with inputs from processing the configuration of PM features

driving the PM accumulation process. More interestingly, we found that this 'PM excitation' was stronger for target PM ships than non-target PM ships. Target ships may induce greater PM excitation because other ship features may provide similar inputs to PM ship features (e.g., the blue smoke stack feature might provide inputs similar to the 'life boat' or 'flag' PM features, because they are both partially blue and in similar locations, see Figure 4). If this is the case, our finding of greater PM excitation for ongoing task targets may not generalize to tasks in which PM-related evidence and ongoing task target evidence is less similar.

We also found evidence of 'reactive inhibition' of ongoing task processing on PM ships. That is, ongoing task accumulation rates were lower for decisions to PM ships than non-PM ships. This effect was strong for both targets and non-targets. We found that inhibition mostly affected the quantity of ongoing task processing, with a smaller effect on the quality of processing, which is very consistent with the 'feedforward' nature of inhibition in PMDC in which inhibition flows forward from the encoding stage, rather than occurring on-line during the decision process. Our correlations across participants revealed that individual differences in reactive inhibition were a large determinant of PM accuracy, replicating our work from basic paradigms (Strickland et al., 2018). Thus, it appears that reactive inhibition of ongoing task processing upon encountering PM events is an important mechanism underlying PM accuracy in the current paradigm.

Multiple Stimulus Encoding. On each trial in our task, participants were presented with three stimuli at a time, and we added a 'response order' factor to our data to capture possible performance differences between earlier and later responses in the trials. We found some effects of response order. We found a decrease in target accuracy (around 2-3%) for the last response on each trial, without an equal effect on non-target accuracy. This effect, although small, is

consistent with the Palada et al. (2018) finding that target accuracy is selectively affected, and can be reduced to chance, by tight deadlines – the third ship on each trial is most likely to be responded to with little time left until trial deadline. The strongest effect of response order by far was an increase in RT for the first ship responded to. This is likely due to orienting attention to the locations of the 3 ships presented, parallel encoding, and orienting the motor response to the first ship. To account for this, we included a different non-decision time for the first trial response than subsequent responses in our model. Replicating Palada et al. (2018), we found that the non-decision time for the first ship was longer than the others.

Four participants responded to PM ships first on the trials that they were presented much more often than second or third on those trials. This suggests that those participants were able to detect PM features at the beginning of the trial, perhaps due to extended encoding. Consistent with this, these four participants had longer non-decision times for the first response than other participants. This illustrates that in complex, multi-stimulus tasks, PM-related adaptive behavior can vary across individuals. More complex paradigms bring the possibility of there being multiple effective strategies (e.g., noticing PM features during encoding vs. noticing PM features during the standard accumulation process), and individuals may differ in adoption of strategies and under what circumstances they apply them. In addition to extended encoding, there is likely initial time devoted to eye saccades, and motor responding. Future work may investigate these further with appropriate tools such as eye tracking and mouse tracking. It might also be possible to experimentally manipulate strategy (e.g., instruct participants it is best to respond to PM ships before ongoing task ships).

Limitations

Before outlining the potential applied implications of our findings, it is critical to acknowledge some potential limitations. We designed our simulated maritime task to be more representative of the conditions operators in complex dynamic work domains face, relative to the previous application of PMDC. Nevertheless, we acknowledge that there are issues associated with generalizing our findings to field operations. For one, the trials in our experiment were discrete – three ships appeared on the camera view at a time and exited at the same time. By contrast, real-world monitoring tasks are continuous, involving multiple stimuli with variable, and often unpredictable, onsets and durations. Under these conditions, patterns of proactive control and capacity allocation may differ. Thus, it would be worthwhile extending PMDC to a more continuous monitoring task. Our paradigm also encouraged speedy serial decision making (i.e., one ship decision at a time), without attention switching before each decision is made. By contrast, in many applied settings, perceptual display items evolve slowly over time, and it is likely operators distribute their attention across stimuli cyclically. It is may be possible to incorporate attention switching into our framework using 'piecewise' accumulation (Holmes, Trueblood & Heathcote, 2016) and eye-tracking measurements, such that evidence accumulation rates about an event can change when gaze moves to another display item.

We are also limited by the fact that our data did not come from expert participants. Whereas our student participants appeared to share resources between PM monitoring and the ongoing task, with practice the performance of expert operators might become more automatic (Logan, 1988), shifting the 'red zone' to a higher level of task demand. Consistent with practice reducing the resource requirements of the task, we observed a smaller PM cost to target detection accuracy on the second session of our experiment. Although the model we reported in detail here

did not estimate ongoing task capacity separately for session one and two (due to concerns about cutting data too finely), we explored such a model, and report the parameter estimates in the supplementary materials. This exploration did indeed indicate more PM cost to ongoing task capacity in session one than in session two. In all other respects it led to similar conclusions to the modeling reported in text. Future work could potentially benefit from formally specifying why these practice effects occur. For example, a model of practice could potentially form a front-end to the PMDC measures of capacity demands.

Practical Implications and Future Directions

Despite these potential limitations, our results have several potential practical implications, and open future directions for research. Our findings regarding capacity and control help to bridge the gap between basic laboratory-based PM research and practical applications of PM theory. For example, we found that proactive and reactive control over ongoing task decision processes support event-based PM, generalizing Strickland et al (2018)'s findings from a simple PM paradigm to a more representative task. This finding may have implications for work design. Both proactive and reactive control mechanisms depend upon delaying ongoing task decision making to facilitate PM retrieval. Interventions may emulate such control. For example, when work designers know that responding to an expected event-based PM event is safety-critical, it may be useful to externally delay operator responses to non-critical stimuli to allow more time for evidence accumulation for the PM action.

Our model indicated that PM demands cost ongoing task processing in the current experiment, whereas models of performance in simple paradigms did not (e.g., Strickland et al., 2018), demonstrating that a lack of capacity sharing between processes in simple laboratory paradigms does not rule out capacity sharing in more complex tasks. Thus, when studying PM

and ongoing task capacity sharing in view of practical applications, it is important to use laboratory tasks with sufficiently representative task demands. This finding also has implications for work design. It suggests that in complex, dynamic tasks, event-based PM could potentially benefit from reducing the capacity burden of ongoing tasks (e.g., via work design inventions such as task automation; Endsley, 2017; Wickens, 2018). Similarly, our findings suggest that for successful performance of both concurrent PM task and ongoing tasks, it is critical to keep workload within safe limits. This may involve planning ahead to reduce the burden of workload with external factors (e.g., by coordinating with other people) (Loft et al. 2007; Neal et al., 2014). Operators may also manage workload adaptively right as they encounter mentally demanding events (Loft et al. 2007; Neal et al., 2014), by adjusting their cognitive operations with mechanisms such as those identified in the current study (shunting capacity between tasks, proactive threshold shifts, reactive inhibition).

A more general line of practical implications follows from our finding that PMDC is a viable model of our complex, dynamic task, and can account for and explain the effects of PM in the 'red zone' of task demand. This suggests that PMDC could potentially serve as a quantitative 'human performance model' (Byrne & Pew, 2009) for many applications of event-based PM. Our modelling does not aim to replace traditional methods such as experimentation, field research, and task analysis. Instead, it can complement these methods, by offering insights into the ecological work contexts that traditional methods indicate there is potentially a high payoff for further understanding. The PMDC architecture is tractable, relatively easy to apply, and offers numerous advantages over verbal theorizing. For example, PMDC could serve as a 'measurement model', similar to how signal detection theory is often used by human factors practitioners (Byrne & Pew, 2009), but using RT as well as accuracy to measure PM processes

(i.e., a “dynamic signal detection theory”). In addition, as PMDC’s accumulation rate parameters are potentially more sensitive to the capacity for information processing than manifest data, PMDC could be used to assess the effectiveness of training or interventions that attempt to reduce cognitive load on the operator. Furthermore, as illustrated in the current paper with our use of ‘posterior exploration’, PMDC allows us to quantify the effects of various model mechanisms. As such, PMDC could be a useful tool to not only identify the latent effects of interventions, but also quantify how much they directly affected a behavioral feature of interest. For example, if an intervention aimed to increase the caution with which operators make ongoing task decisions, PMDC could directly measure whether the intervention succeeded in doing so, and quantify the benefits of any caution increases to PM performance. We do acknowledge that applying PMDC as a ‘measurement model’ requires observing adequate numbers of PM responses to estimate parameters, which could be challenging in some applied contexts, but this could be improved with a hierarchical model of the population parameters (observing many humans, rather than many responses per human), and the incorporation of prior information to constrain parameter estimates.

As a quantitative human performance model, PMDC may also serve to unify disparate human data beyond RT and response choice. For example, it is straightforward to explore correlations between model parameters and individuals abilities in a fully Bayesian way using posterior correlations (Ly, Boehm, et al., 2018). This could help to identify the measures (e.g., working memory scores) that are most associated with causes of PM performance in complex, dynamic tasks. It is also possible to include trial level covariates in the model, in which case model parameters are specified as a function of some other measurement (e.g., biometrics such as pupil dilation) around the time of the decision. With this approach, we could find the

measurements that best correspond to PM decision parameters. It might even be possible to incorporate data from previous responses to determine the parameters of the current PM decision, providing a model-based method to predict PM failures before they happen.

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? *Journal of Experimental Psychology: Learning, Memory, and Cognition, Advance online publication*. https://doi.org/10.1037/xlm0000504

Ball, B. H., & Aschenbrenner, A. J. (2018). The importance of age-related differences in prospective memory: Evidence from diffusion model analyses. *Psychonomic Bulletin & Review, 25*(3), 1114–1122. https://doi.org/10.3758/s13423-017-1318-4

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Boywitt, C. D., & Rummel, J. (2012). A diffusion model analysis of task interference effects in prospective memory. *Memory & Cognition, 40*(1), 70–82. https://doi.org/10.3758/s13421-011-0128-6

Braver, T. S. (2012). The variable nature of cognitive control: A dual mechanisms framework. *Trends in Cognitive Sciences, 16*(2), 106–113. https://doi.org/j.tics.2011.12.010

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Bundesen, C. (1990). A theory of visual attention. *Psychological Review, 97*(4), 523–547. https://doi.org/10.1037/0033-295x.97.4.523

Byrne, M. D., & Pew, R. W. (2009). A history and primer of human performance modeling. *Reviews of human factors and ergonomics, 5*(1), 225-263.

Dismukes, R. K., Berman, B. A., & Loukopoulos, L. D. (2007). *The limits of expertise*. Hampshire, England: Ashgate Publishing Limited.

Dismukes, R. K. (2012). Prospective memory in workplace and everyday situations. *Current Directions in Psychological Science*, *21*(4), 215–220. https://doi.org/10.1177/0963721412447621

Donkin, C., Brown, S. D., & Heathcote, A. (2009). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review*, *16*(6), 1129–1135. https://doi.org/10.3758/PBR.16.6.1129

Donkin, C., Little, D. R., & Houpt, J. W. (2014). Assessing the speed-accuracy trade-off effect on the capacity of information processing. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1183–1202. https://doi.org/10.1037/a0035947

Eidels, A., Donkin, C., Brown, S. D., & Heathcote, A. (2010). Converging measures of workload capacity. *Psychonomic Bulletin & Review*, *17*(6), 763–771. https://doi.org/10.3758/PBR.17.6.763

Einstein, G. O., & McDaniel, M. A. (1990). Normal aging and prospective memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(4), 717–726. https://doi.org/10.1037/0278-7393.16.4.717

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. *Journal of Experimental Psychology: General*, *134*, 327–342. https://doi.org/10.1037/0096-3445.134.3.327

Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned from Human-Automation Research. *Human Factors, 59(*1), 5–27. https://doi.org/10.1177/0018720816681350

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis, Third Edition.* Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Gobell, J. L., Tseng, C., & Sperling, G. (2004). The spatial distribution of visual attention. *Vision Research*, *44*(12), 1273–1296. https://doi.org/10.1016/j.visres.2004.01.012

Hart, S., & Wickens, C. (2010). Cognitive workload. In *NASA Human Systems Integration Handbook* (pp. 190–222). Washington, DC: NASA.

Heathcote, A., Lin, Y.-S., Reynolds, A., Strickland, L., Gretton, M., & Matzke, D. (in press). Dynamic models of choice. *Behavior Research Methods*.

Heathcote, A., Loft, S., & Remington, R. W. (2015). Slow down and remember to remember! A delay theory of prospective memory costs. *Psychological Review, 122*(2), 376–410. https://doi.org/10.1037/a0038952

Holmes, W. R., Trueblood, J. S., & Heathcote, A. (2016). A new framework for modeling decisions about changing information: The Piecewise Linear Ballistic Accumulator model. *Cognitive psychology, 85*, 1-29.

Horn, S. S., & Bayen, U. J. (2015). Modeling criterion shifts and target checking in prospective memory monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 95–117. https://doi.org/10.1037/a0037676

Horn, S. S., Bayen, U. J., & Smith, R. E. (2011). What can the diffusion model tell us about prospective memory? *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *65*, 69–75. https://doi.org/10.1037/a0022808

Horn, S. S., Bayen, U. J., & Smith, R. E. (2013). Adult age differences in interference from a prospective-memory task: a diffusion model analysis. *Psychonomic Bulletin & Review*, *20*, 1266 –1273. https://doi.org/10.3758/s13423-013-0451-y

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Lerche, V., & Voss, A. (2017). Experimental validation of the diffusion model based on a slow response time paradigm. *Psychological Research*, 1–16.

Loft, S. (2014). Applying Psychological Science to Examine Prospective Memory in Simulated Air Traffic Control. *Current Directions in Psychological Science*, *23*(5), 326–331. https://doi.org/10.1177/0963721414545214

Loft, S., Finnerty, D., & Remington, R. W. (2011). Using spatial context to support prospective memory in simulated air traffic control. *Human Factors*, *53*(6), 662–671.

Loft, S., & Remington, R. W. (2010). Prospective memory and task interference in a continuous monitoring dynamic display task. *Journal of Experimental Psychology: Applied*, *16*, 145. https://doi.org/10.1037/a0018900

Loft, S., & Remington, R. W. (2013). Wait a second: Brief delays in responding reduce focality effects in event-based prospective memory. *The Quarterly Journal of Experimental Psychology*, *66*(7), 1432–1447. https://doi.org/10.1080/17470218.2012.750677

Loft, S., Sanderson, P., Neal, A., & Mooij, M. (2007). Modeling and predicting mental workload in en route air traffic control: Critical review and broader implications. *Human Factors*, *49*(3), 376–399.

Loft, S., Smith, R. E., & Bhaskara, A. (2011). Prospective memory in an air traffic control simulation: External aids that signal when to act. *Journal of Experimental Psychology: Applied, 17*(1), 60.

Loft, S., Smith, R. E., & Remington, R. W. (2013). Minimizing the disruptive effects of prospective memory in simulated air traffic control. *Journal of Experimental Psychology: Applied, 19*(3), 254.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological review, 95*(4), 492-527. https://doi.org/10.1037/0033-295X.95.4.492

Logan, G. D., Van Zandt, T., Verbruggen, F., & Wagenmakers, E.-J. (2014). On the ability to inhibit thought and action: General and special theories of an act of control. *Psychological Review, 121*(1), 66–95. https://doi.org/10.1037/e633262013-227

Lourenço, J. S., White, K., & Maylor, E. A. (2013). Target context specification can reduce costs in nonfocal prospective memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1757–1764. https://doi.org/10.1037/a0033702

Ly, A., Boehm, U., Heathcote, A., Turner, B. M., Forstmann, B., Marsman, M., & Matzke, D. (2018). A flexible and efficient hierarchical bayesian approach to the exploration of individual differences in cognitive-model-based neuroscience. In A. A. Moustafa (Ed.), *Computational models of brain and behavior* (pp. 467–480). London, UK: Wiley Blackwell.

Ly, A., Marsman, M., & Wagenmakers, E. (2018). Analytic posteriors for Pearson's correlation coefficient. *Statistica Neerlandica*, *72*(1), 4–13.

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. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *29*, 861–870. https://doi.org/10.1037/0278-7393.29.5.861

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, 1142–1160.

Morrow, D. (2018). Publishing papers that matter. *Journal of Experimental Psychology: Applied*, *24*(1), 1–2. https://doi.org/10.1037/xap0000141

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review*, *86*(3), 214–255. https://doi.org/10.1037/0033-295x.86.3.214

Neal, A., Hannah, S., Sanderson, P., Bolland, S., Mooij, M., & Murphy, S. (2014). Development and validation of a multilevel model for predicting workload under routine and nonroutine conditions in an air traffic management center. *Human factors*, *56*(2), 287-305.

Nilsson, M., Van Laere, J., Ziemke, T., & Edlund, J. (2008). Extracting rules from expert operators to support situation awareness in maritime surveillance. In *Information Fusion, 2008 11th International Conference on* (pp. 1–8). IEEE.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, *7*(1), 44–64. https://doi.org/10.1016/0010-0285(75)90004-3

Palada, H., Neal, A., Tay, R., & Heathcote, A. (2018). Understanding the causes of adapting, and failing to adapt, to time pressure in a complex multi-stimulus environment. *Journal of Experimental Psychology: Applied*. https://doi.org/10.1037/xap0000176

Palada, H., Neal, A., Vuckovic, A., Martin, R., Samuels, K., & Heathcote, A. (2016). Evidence accumulation in a complex task: Making choices about concurrent multiattribute stimuli under time pressure. *Journal of Experimental Psychology: Applied*, *22*(1), 1.

Pinheiro, J. C., & Bates, D. M. (2000). Linear mixed-effects models: basic concepts and examples. *Mixed-Effects Models in S and S-Plus*, 3–56.

R Development Core Team. (2018). *R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Austria*. Retrieved from https://www.R-project.org/

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

Rothschild, J. M., Landrigan, C. P., Cronin, J. W., Kaushal, R., Lockley, S. W., Burdick, E., ... & Bates, D. W. (2005). The Critical Care Safety Study: The incidence and nature of adverse events and serious medical errors in intensive care. *Critical care medicine, 33*(8), 1694-1700.

Rummel, J., Smeekens, B. A., & Kane, M. J. (2017). Dealing with prospective memory demands while performing an ongoing task: Shared processing, increased on-task focus, or both? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(7), 1047–1062.

Shorrock, S. T. (2005). Errors of memory in air traffic control. *Safety Science*, *43*(8), 571–588. https://doi.org/10.1016/j.ssci.2005.04.001

Smith, R. E. (2003). The cost of remembering to remember in event-based prospective memory: Investigating the capacity demands of delayed intention performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 347–361. https://doi.org/10.1037/0278- 7393.29.3.347

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Stokes, D. E. (2011). *Pasteur's quadrant: Basic science and technological innovation.* Washington, D.C: Brookings Institution Press.

Strickland, L., Heathcote, A., Remington, R. W., & Loft, S. (2017). Accumulating evidence about what prospective memory costs actually reveal. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(10), 1616.

Strickland, L., Loft, S., Remington, R. W., & Heathcote, A. (2018). Racing to remember: A theory of decision control in event-based prospective memory. *Psychological Review*. https://doi.org/10.1037/rev0000113

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*(3), 368.

Wickens, C. D. (1980). The structure of attentional resources. In R. Nickerson (Ed.), *Attention and performance VIII* (pp. 239–257). Hillsdale, NJ: Erlbaum.

Wickens, C. D. (2018). Automation Stages & Levels, 20 Years After. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 35–41. https://doi.org/10.1177/1555343417727438

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2015). Mental workload, stress, and individual differences: cognitive and neuroergonomic perspectives. *Engineering Psychology and Human Performance (International Edition)*, 346–376.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: mental workload in ergonomics. *Ergonomics, 58*(1), 1–17.

Table 1

Experimental counterbalance. Note that we have marked control to indicate one control block, whereas PM indicates two PM blocks, one after the other. We collected more PM blocks than control in order to collect more observations of PM responses for modeling.

BLOCK ORDER	PM RESPONSE BUTTON
1: control-PM-PM-control	RED
2: control-PM-PM-control	GREEN
3: PM-control-control-PM	RED
4: PM-control-control-PM	GREEN

Table 2

Correct-response latencies (RT, in seconds) and accuracy (Acc, as a proportion) for ongoing task (feature classification) performance for target stimuli (4+ features) and non-target stimuli (<4 features) in control and PM blocks.

		Control				PM			
Session	Stimulus	RT	SE	Acc	SE	RT	SE	Acc	SE
1	Target	1.80	0.012	0.94	0.004	2.02	0.010	0.87	0.004
	Non-Target	2.02	0.013	0.94	0.003	2.03	0.010	0.93	0.003
2	Target	1.77	0.013	0.92	0.004	1.89	0.010	0.91	0.003
	Non-Target	1.88	0.013	0.96	0.003	1.95	0.010	0.95	0.002

Table 3

Correct-response latencies and accuracy for PM task performance (two specific features present) for target stimuli (4+ features) and non-target stimuli (< 4 features) in control and PM blocks.

	Session 1				Session 2			
Stimulus	RT (s)	SE	Acc	SE	RT (s)	SE	Acc	SE
Target	1.89	0.012	0.82	0.012	1.73	0.012	0.77	0.012
Non-Target	1.92	0.013	0.66	0.013	1.71	0.013	0.66	0.013

Table 4

Correct-response latencies and accuracy for ongoing task (feature classification) performance by response order in control and PM blocks.

			Control				PM			
Session	Task	Response Order	RT (s)	SE	Acc	SE	RT (s)	SE	Acc	SE
1	Ongoing Task	1	2.47	0.018	0.96	0.004	2.65	0.013	0.91	0.004
		2	1.60	0.012	0.95	0.004	1.73	0.010	0.90	0.004
		3	1.64	0.012	0.93	0.004	1.70	0.010	0.89	0.004
	PM Task	1					2.46	0.037	0.76	0.037
		2					1.53	0.029	0.75	0.029
		3					1.50	0.027	0.72	0.027
2	Ongoing Task	1	2.41	0.018	0.96	0.004	2.53	0.013	0.94	0.003
		2	1.52	0.011	0.94	0.004	1.61	0.010	0.94	0.003
		3	1.53	0.011	0.93	0.005	1.62	0.010	0.92	0.004
	PM Task	1					2.31	0.037	0.72	0.037
		2					1.41	0.027	0.72	0.027
		3					1.38	0.02	0.70	0.023

Table 5

Priors for the parameters of the LBA models fitted to our data.

Model Parameter	Distribution	M	SD	Lower	Upper
A	Truncated Normal	1	1	0	10
B	Truncated Normal	2	1	0	None
*v (*Ongoing match)	Truncated Normal	1	2	None	None
*v (*Ongoing mismatch)	Truncated Normal	0	2	None	None
*v (*PM match)	Truncated Normal	1	2	None	None
v (PM false alarm)	Truncated Normal	-1	2	None	None
s_v	Truncated Normal	1	1	0	None
t0	Uniform			0.1	3

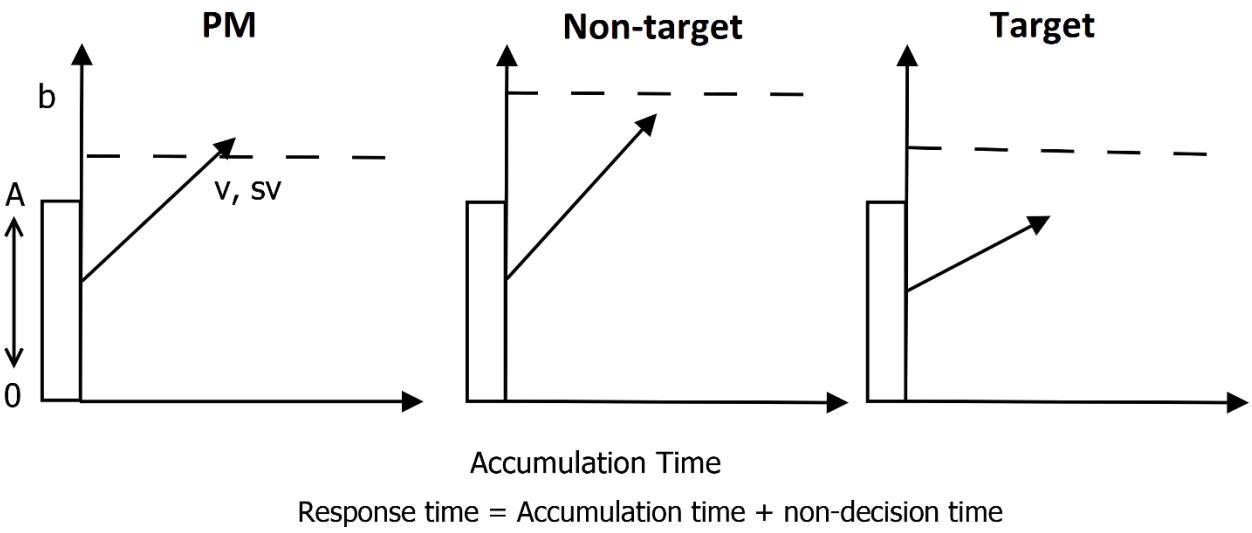

Figure 1. PMDC as it applies to our target detection task. Evidence for each response is initially drawn from a uniform distribution on the interval [0, *A*]. Over time, evidence accumulates towards each response at rates drawn from normal distributions with mean *v*, and standard deviation *sv*. The first accumulator to reach its threshold, *b*, determines the overt response. Total RT is given by total accumulation time plus non-decision time. Note that we depict a bias against non-target responding, as may be the case in classification tasks where missing a target (e.g., failing to identify a target ship) is often more costly than a false alarm (e.g., flagging a non-target ship for further investigation).

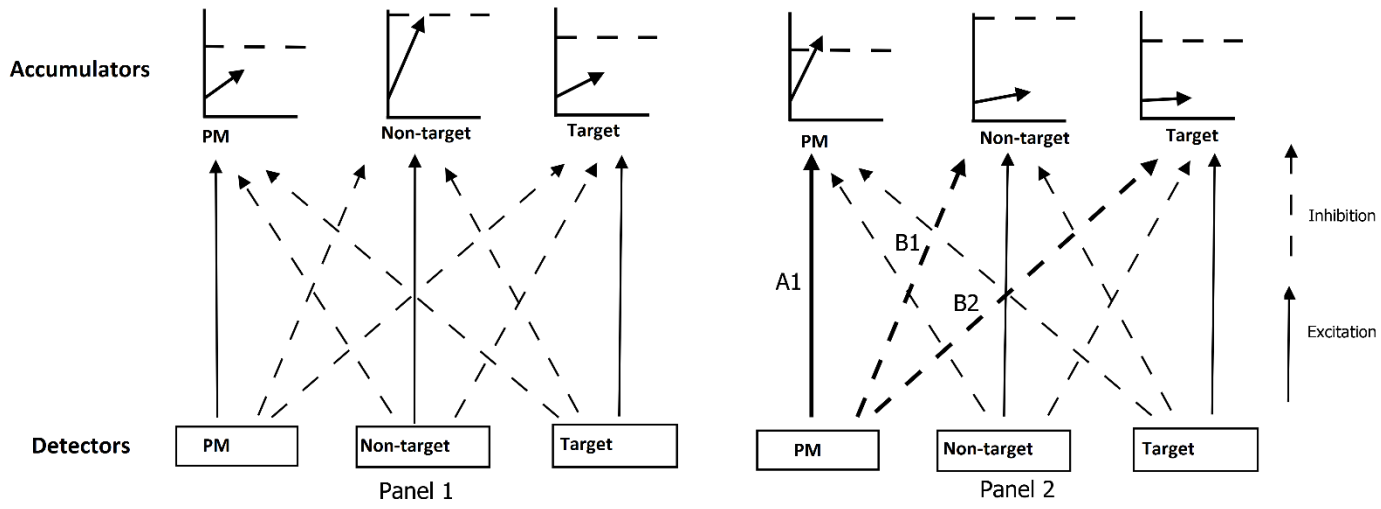

Figure 2. The PMDC architecture's proposed reactive control structure, as it would apply to our target detection task with a PM requirement. Panel 1 depicts the basic structure. There is a detector for each possible decision. PM stimulus inputs may activate the PM detector. Panel 2 illustrates the possible effects of PM activation on the decision process. Activation may excite the corresponding response, increasing accumulation (e.g., A1), but also inhibit competing responses, slowing their accumulation (e.g., B1 & B2).

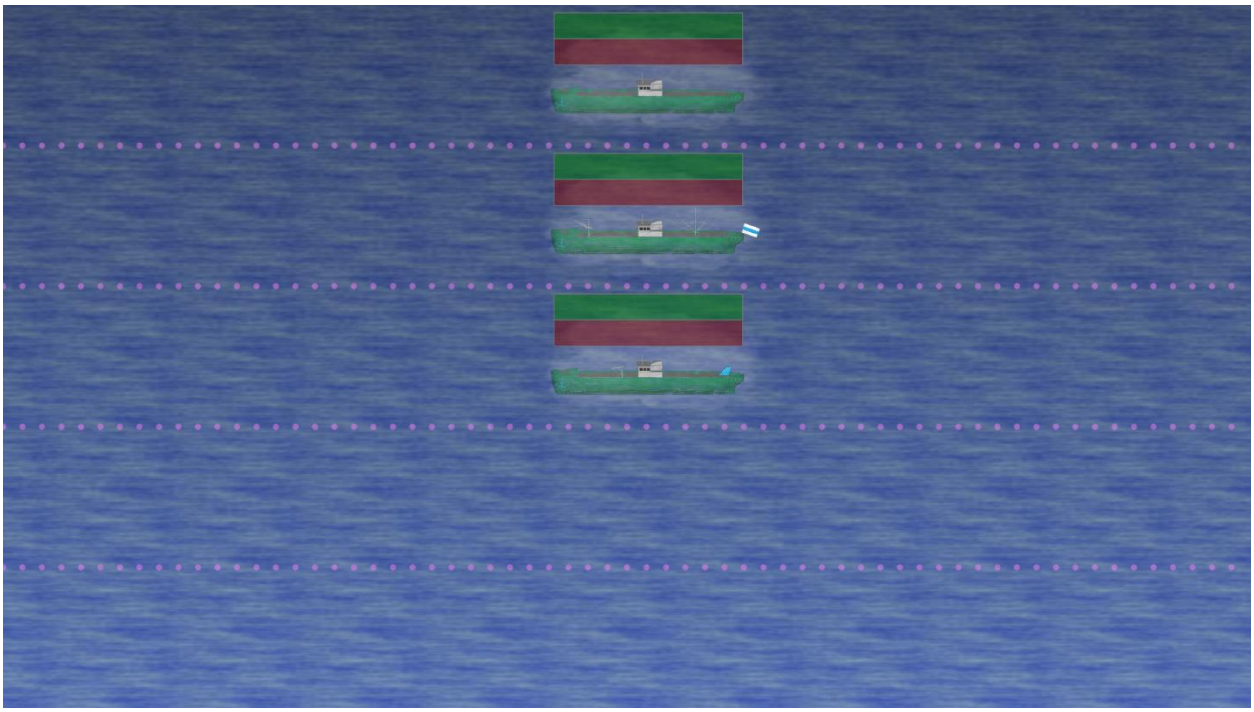

Figure 3. Screenshot of the UAV task, with possible non-target (red box) and target (green box) responses required to each ship. Note that the noise and fog overlay are reduced for illustrative purposes.

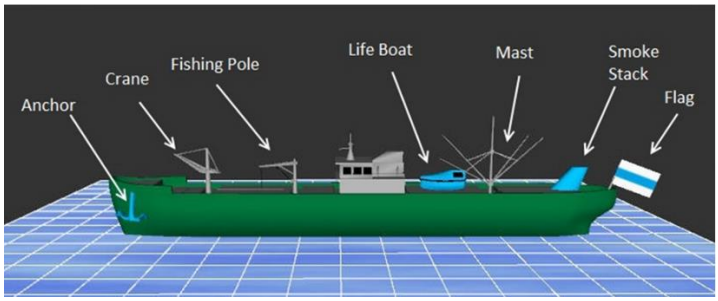

Figure 4. Schematic of ship stimulus displaying all seven possible features.

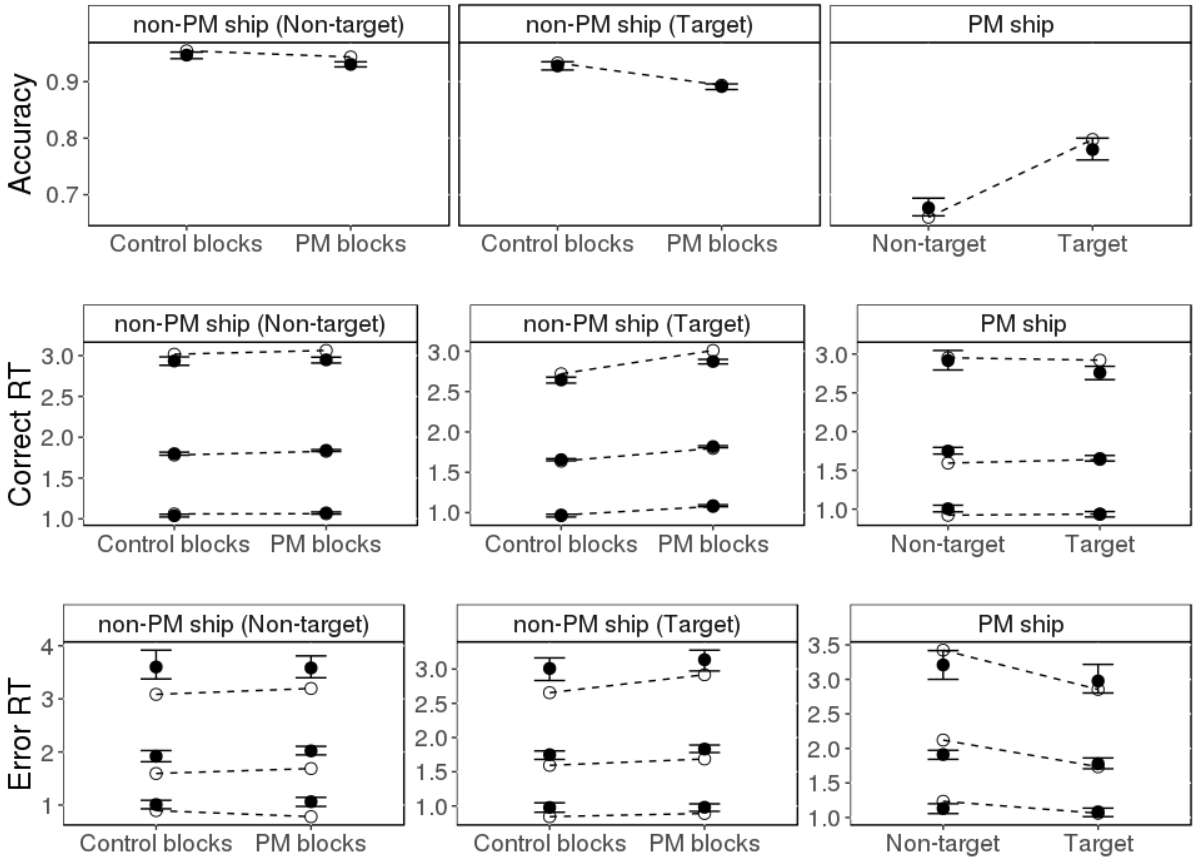

Figure 5. Fits of our chosen (top model) to the data. The data are illustrated by the white dots. The model predictions are the black dots. The error bars around the predictions are the 95% credible intervals of the predictions (i.e., the uncertainty). The RT graphs depict quantiles, including the median (the middle values), the 0.1 quantile (fastest 10% of responses), and the 0.9 quantiles (the slowest 10% of response times). We plot fits to the participant-averaged data, which we obtained by concatenating all data together, and then summarising, for both the data and the model. The PM error RTs depicted correspond to the correct ongoing task decision on PM ships (e.g., target decision to a target PM ship). The plots do not depict incorrect ongoing task RTs to PM ships, nor PM RTs to non-PM ships, because they were rarely observed.

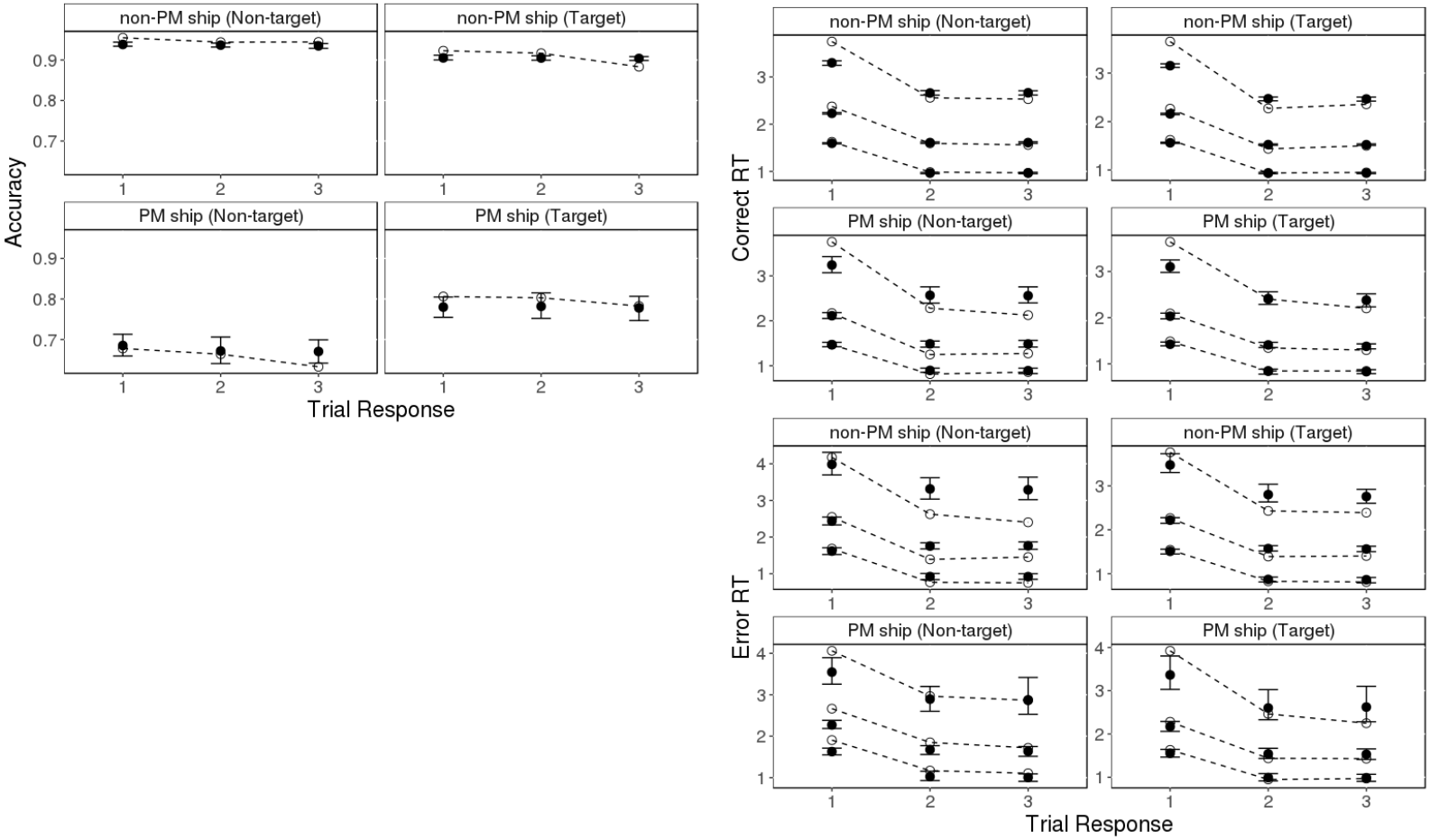

Figure 6. Fits of our chosen (top model) to the data by response order. The data are illustrated by the white dots. The model predictions are the black dots. The error bars around the predictions are the 95% credible intervals of the predictions (i.e., the uncertainty). The RT graphs depict quantiles, including the median (the middle values), the 0.1 quantile (fastest 10% of responses), and the 0.9 quantiles (the slowest 10% of response times). We plot fits to the participant-averaged data, which we obtained by concatenating all data together, and then summarising, for both the data and the model. The PM error RTs depicted correspond to the correct ongoing task decision on PM ships (e.g., target decision to a target PM ship). The plots do not depict incorrect ongoing task RTs to PM ships, nor PM RTs to non-PM ships, because they were rarely observed.

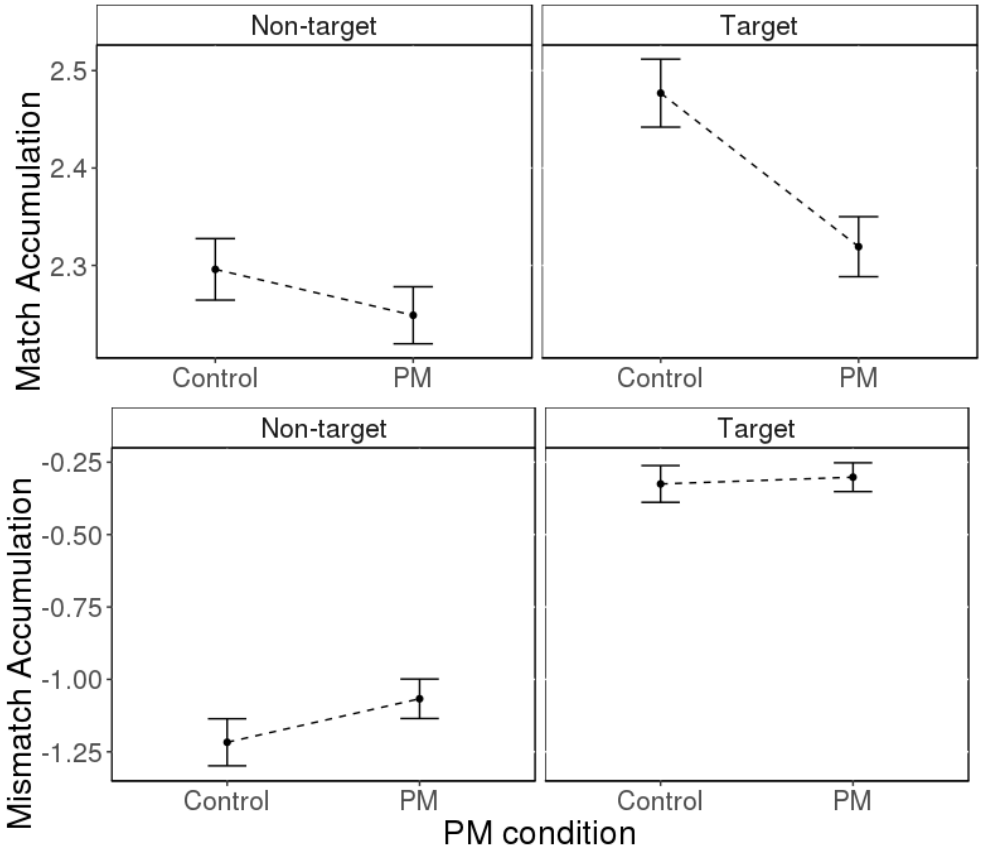

Figure 7. The estimated accumulation rates to non-PM ships. The dots indicate the posterior means of the subject-average parameters, and the error bars are the mean plus or minus the standard deviations of the posteriors. The x axis indicates condition (control vs PM). The panels indicate stimulus type (non-target vs target). The top panels illustrate accumulation to 'match' accumulation rates (that is the accumulation towards the correct response, e.g., non-target accumulation to non-targets). The bottom panels illustrate ongoing task 'mismatch' accumulation (e.g., target accumulation to non-targets).

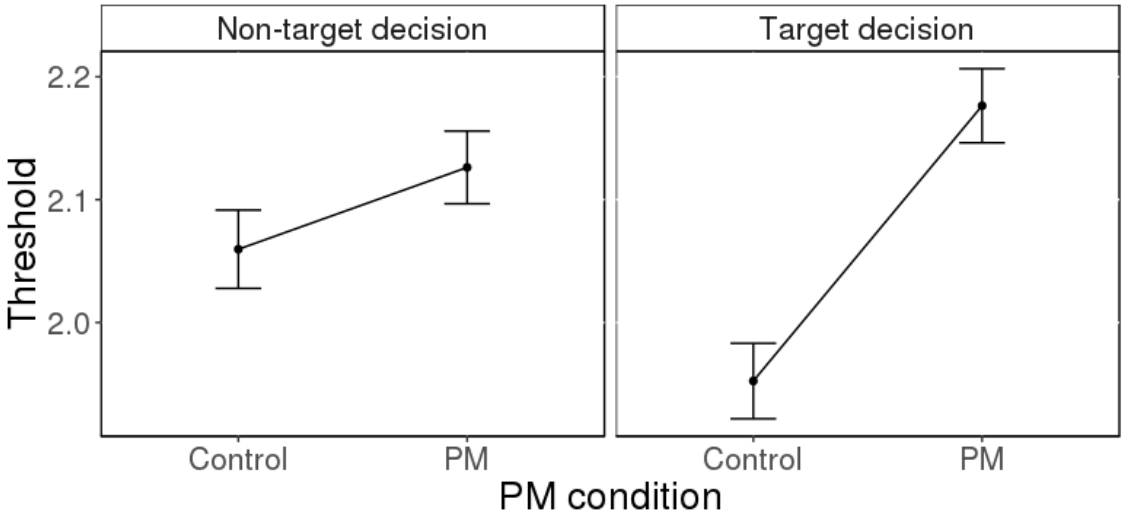

Figure 8. The estimated thresholds to ongoing task decisions. The dots indicate the posterior means of the subject-average parameters, and the error bars are the means plus or minus the standard deviations of the posteriors. The x axis indicates condition (control vs PM). The panels indicate the latent response accumulator (non-target vs target).

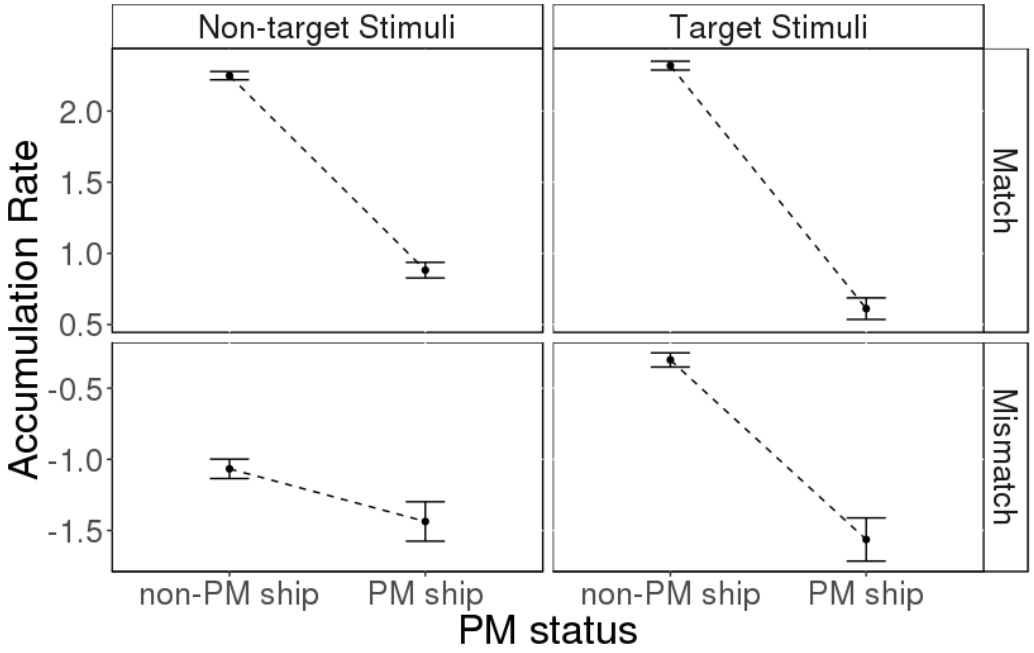

Figure 9. Ongoing task accumulation rates to non-PM ships compared with PM ships. The difference in rates between non-PM and PM ships indicates reactive inhibitory control. The dots indicate the posterior means of the subject-average parameters, and the error bars are the posterior means plus or minus the standard deviations of the posteriors. The x axis indicates whether the ship was a PM ship or not. The panels indicate stimulus type in terms of the ongoing task rule (non-target vs target). The top panels illustrate accumulation to 'correct' accumulation rates in terms of the ongoing task (that is the accumulation towards the correct response, e.g., non-target accumulation to non-targets). The bottom panels illustrate ongoing task 'error' accumulation (e.g., target accumulation to non-targets).

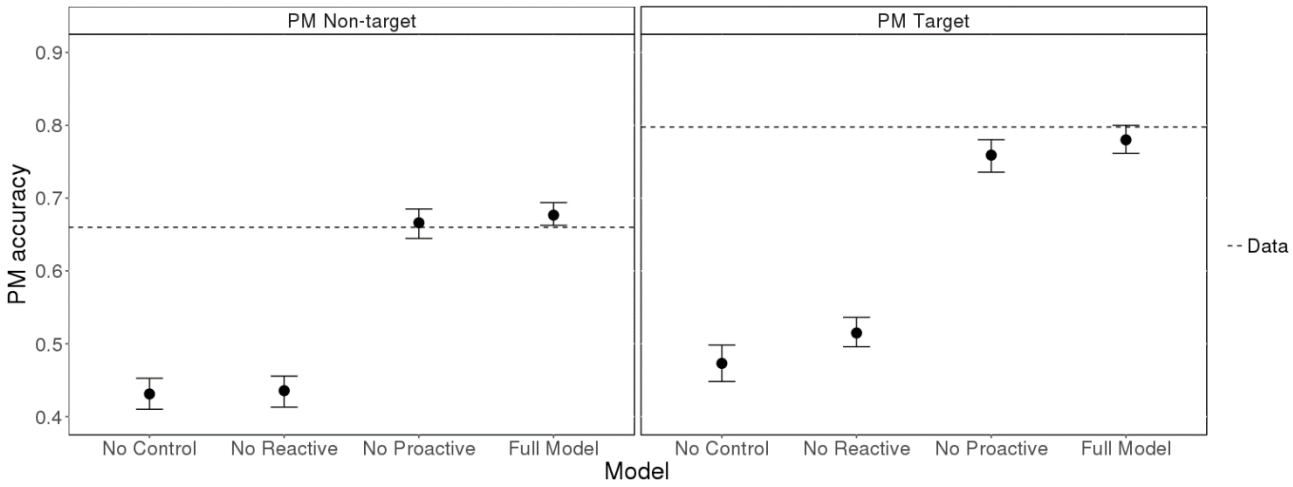

Figure 10. Posterior exploration of the role of model mechanisms underlying PM accuracy. The dashed line indicates the observed PM accuracy in the data. The black dots indicate the posterior mean predicted by each model, and the error bars are the 95% credible intervals. The full model is that reported in text, whereas the other models were obtained by removing effects from the full model after parameter estimation. These graphs depict fits to the participant average data. This was obtained by concatenating all data together, and then summarizing, for both the data and the model predictions.

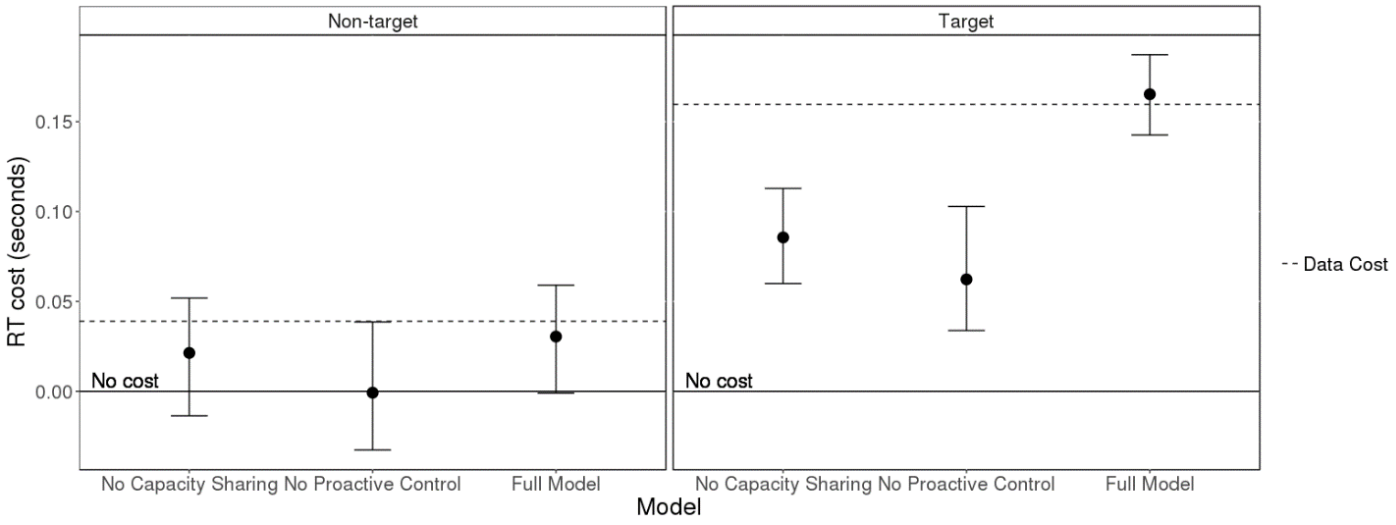

Figure 11. Posterior exploration of the role of model mechanisms in PM cost. The dashed line indicates the observed PM cost in the data. The black dots indicate the posterior mean predicted by each model, and the error bars are the 95% credible intervals. The full model is that reported in text, whereas the other models were obtained by removing effects from the full model after parameter estimation. These graphs depict fits to the participant average data. This was obtained by concatenating all data together, and then summarizing, for both the data and the model predictions.

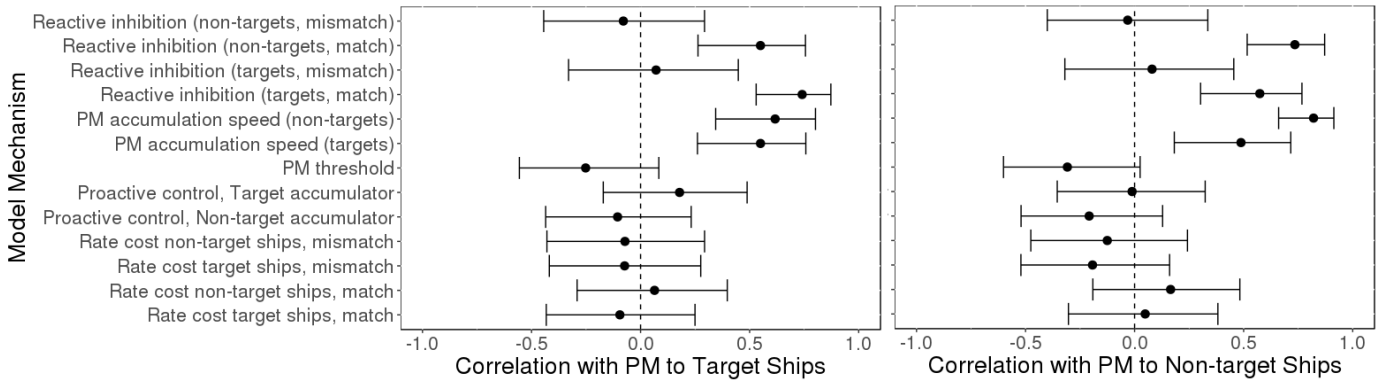

Figure 12. Posterior correlations of model mechanisms with PM accuracy. The distributions were obtained by separately correlating model mechanisms and PM accuracy across participants for each posterior sample. The correlations were then corrected for population level inference (Ly, Marsman, & Wagenmakers, 2018). The black dots indicate the means of the posteriors, and the bars are the 95% credible intervals. The dashed line sits on a correlation value of 0.