Introduction and Method for ATC Experiment E1

R. Boag

Introduction

For the past 50 years, much psychological research on decision-making has used two-alternative forced-choice (2AFC) tasks to make theoretical inferences about the latent cognitive processes that drive performance (Laming, 1968; Link & Heath, 1975; McClelland, 1979; Stone, 1960; for reviews see Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Ratliff & Smith, 2004). In a typical 2AFC task, decision-makers are presented with a series of stimuli about which a decision or response must be made (e.g., is this letter string a word?; is this arrow pointing left or right?). Conventionally, performance on these kinds of simple decision-making tasks had been analysed by comparing mean response time (RT), mean accuracy, or receiver operating characteristic (ROC) curves (Swets, 1973) between experimental groups, or by using methods related to signal detection theory (SDT; Green & Swets, 1974).

More recently, theoretical advances in decision-making research have stemmed from the application of accumulate-to-threshold models (Forstmann, Ratcliff, & Wagenmakers, 2016). Accumulate-to-threshold models are a class of computational process model which formalise the latent cognitive processes theorised to underlie decision-making. When fit to data, such models provide a full quantitative account of both RT distributions and the accuracy of decisions. The central feature of accumulate-to-threshold models is that decision-making is conceptualised as a process of taking repeated samples of information from the environment until enough evidence has been obtained to trigger a response or action.

This study aims to address this gap by extending the LBA model to a complex dynamic ATC task representative of a broad set of applied tasks common to many work settings. We aim to evaluate the impact of two of the most common external enivironmental variables that commonly impact tasks in applied settings: time pressure, and prospective memory (PM) demands. Finally, we

demonstrate the usefulness of our modelling approach in illuminating the latent cognitive processes that drive task performance and how such processes change in response to different environmental conditions (time pressure and PM demand). Thus showing the utility of this approach in answering both theoretical and applied questions. The following section outlines some potential issues with applying a model of simple choice such as the LBA to a more complex and dynamic task.

Although there are many such models, which can differ on numerous dimensions (e.g., linear vs nonlinear accumulation; independent vs dependent accumulation, fixed-rate vs decaying accumulation, static vs collapsing response thresholds, and so on; Hawkins, Forstmann, Wagenmakers, Ratcliff, Brown, 2015; Smith, 2010; Usher & McClelland, 2001; Vickers, 1970), models typically give convergent theoretical interpretations when fit to the same data (Donkin, Brown, Heathcote, & Wagenmakers, 2011; Heathcote & Hayes, 2012).

Accumulate-to-Threshold Models for Basic Tasks

One of the most influentual and widely applied accumulate-to-threshold models for basic tasks is the Linear Ballistic Accumulator (LBA; Brown & Heathcote, 2008). The LBA formalises decision-making as a process of evidence accumulation among independent racing accumulators. Evidence is sampled over time in a linear accumulation process until the evidence in one accumulator reaches a predetermined response boundary. The first accumulator to reach threshold determines the overt response. The LBA models tasks in which a rapid choice between two or more response alternatives must be made, such as brightness discrimination or lexical decision tasks. The core components of the LBA are: the speed or quality of evidence processing, which is quantified by the accumulation rate parameter; the response criterion or threshold, quantified by the height of an accumulator's response boundary; and the nondecision time parameter, which quantifies aspects of processing that fall outside of the accumulation process such as stimulus encoding and motor response time. Several other parameters represent variability in the mean rate and starting point of the accumulation process, and a bias parameter is sometimes used to quantify response biases between different

alternatives. The LBA is easily extended to tasks involving more than two response alternatives, and the absence of non-linear diffusion makes it more computationally tractable. Each LBA accumulator has its own response boundary or threshold, nondecision time, accumulation rate starting point and variabilty parameters. By evaluating which parameters can remain fixed and which must be allowed to vary over experimental conditions, theoretical inferences can be drawn about the latent cognitive processes that underlie decision-making in various tasks and under various conditions (Donkin, Brown, & Heathcote, 2011).

The LBA and DM have yielded important theoretical insights in a wide range of basic and applied domains, including attention and working memory (Sewell, Lilburn, & Smith, 2016), recognition memory (Ratcliff, 1978), workload capacity (Eidels, Donkin, Brown, & Heathcote, 2010), bilingualism (Ong, Sewell, Weekes, & McKague, 2017), sleep deprivation (Ratcliff & van Dongen, 2011), alcohol and drug use (van Ravenzwaaij, Dutilh, & Wagenmakers, 2012), clinical disorders (Moustafa, Kéri, Somlai, Balsdon, Frydecka, Misiak, & White, 2015; White, Ratcliff, Vasey, & McKoon, 2010), consumer choice (Hawkins, Marley, & Heathcote, 2014; Trueblood, Brown, & Heathcote, 2014), and applied aviation and defense research (Palada, Neal, Vuckovic, Martin, Samuels, & Heathcote, 2016; Vuckovic, Kwantes, Humphreys, & Neal, 2014).

Accumulate-to-Threshold Models for Complex Dynamic Tasks

However, most tasks analysed with these models have been basic cognitive or psychophisiological tasks (e.g., perceptual discrimination, recognition) in which decisions typically unfold over very short timescales (i.e., less than 1 to 2 seconds) and with perceptually simple, static stimuli (e.g., gabor patches, letter strings, or simple geomteric shapes). Although such tasks and the associated modelling results have provided many important insights into cognitive and perceptual processes, these tasks have limited generalisability to more applied settings in which decisions often unfold over much longer timescales, stimuli are often more complex and change dynamically over time, and where external environmental variables such as time pressure, task load, and addtional memory

demands often dictate how an operator must approach a given task (Loft, 2014). Given the growing applied and human factors interest in formal models of human performance (Byrne & Gray, 2003; Schweickert, Fisher, & Proctor, 2003), an important goal is the development of models that properly characterise the cognitive processes involved in complex applied tasks and can provide accurate predictions regarding performance under different circumstances (Dismukes, 2008).

In the field of air-traffic control (ATC), for example, controllers are responsible for ensuring the safe and efficient passage of aircraft through a sector of airspace (Durso & Manning, 2009). To this end, controllers must continuously monitor multiple aircraft, which enter and exit the display at different times, and which vary on a number of critical spatial variables such as airspeed, flight level (altitude), climb rate, and heading. In addition, controllers are often required to monitor changing weather conditions, respond to pilot requests, and ensure aircraft maintain adequate lateral and vertical separation (Durso & Manning, 2009). This degree of task complexity is representative of that faced by operators in a wide variety of applied, safety-critical settings, including submarine track management (Loft, Bowden, Braithwaite, Morrell, Huf, & Durso, 2015) and the operation of unmanned aerial vehicles (UAVs; Palada, Neal, Vuckovic, Martin, Samuels, & Heathcote, 2016).

Although these tasks share some similarities with basic lab paradigms (e.g., both UAV and perceptual vigilance tasks involve detecting targets), their level of complexity is clearly very different, which may have implications for our ability to create models of these kinds of tasks. As such, an important research goal is to establish whether such models can be applied to more complex, dynamic, 'real-world' tasks such as ATC in which operators must deal with frequent changes in both display stimuli and aspects of the task environment (Loft, 2014). Importantly, if such tasks can be modelled appropriately, these models may be able to provide valuable insights regarding how people perform these tasks, the strategies they use, and what cognitive processes that drive changes in performance. In addition, insights from model-based analyses be used to inform applied and human factors research regarding practical issues such as task design, training, and the development of decision aids (Dismukes, 2008).

Assumptions of Evidence Accumulation Models and Issues with Modelling Complex Tasks

One reason models of simple choice like the DM and LBA have not seen wide application to more complex decision-making tasks is that the architecture of the models contain certain assumptions about the nature of the cognitive processes driving each decision - assumptions which must be met in order for the models to fit data and provide valid theoretical interpretations (Lerche & Voss, 2017; Ratcliff, 2002; Voss, Nagler, & Lerche, 2013). One assumption with implications for applied tasks is that models assume decisions are the result of a relatively 'pure' evidence accumulation process (i.e., a single continuous stage of information integration; Smith, 2000; Vanderkerckhove & Tuerlinckx, 2007) in which responses are made very quickly (usually within 1 second). This differs from many real-world settings where multiple sources of information must be integrated to form a decison, and where decisions often unfold over much longer timescales (e.g., 2-10 seconds). This difference is important for modelling because long decisions are less likely to be the result of a continuous single stage of uninterrupted sampling and are more likely to be contaminated by other processes such as double-checking, distraction and mind-wandering, interruptions, and attention-switching which are not accounted for in standard simple choice models. Relatedly, the models typically assume that only a single stimulus or source of evidence is considered at a time. This differs from many applied tasks which can have multiple relevant stimuli on display at the same time - all of which could contain decision-critical information and which therefore must be processed simultaneously.

Perhaps most important for complex applied tasks is assumption that information is processed in parallel rather than sequentially (Brown & Heathcote, 2008). This assumption is critical for fitting the models to data and is also arguably the most likely assumption to break down in tasks involving multiple dynamic stimuli. In air-traffic control tasks, for example, operators may scan sequentially between stimuli as well as between other on-screen items (e.g., timers, scale markers, bearing vectors) multiple times before making a response. These kinds of serial decision-making strategies produce RT distributions with very different shapes than those produced by parallel decision-making strategies (Palada, et al., 2016), which may result models being unable to fit data (Ratcliff, 2002; Heathcote, Wagenmakers, & Brown, 2014). In other words, complex, multi-stimulus tasks may encourage different strategies (e.g., serial processing, partial processing, double-checking,

etc.), rely on different cognitive mechanisms (e.g., response competition or interference, nonlinear or piecewise accumulation, evidence decay, urgency signals/collapsing thresholds, etc.), or induce other contaminating factors (e.g., distractions, interruptions, mind-wandering, attention-switching) than those involved in simpler decision-making tasks. At a theoretical level, any model that does not include those processes will be a poor descriptive model of the task and will likely provide poor fits to empirical data.

There have been several recent attempts to model more complex applied tasks with evidence accumulation models like the DM and LBA (Diederich, 1997; Eidels, Donkin, Brown, & Heathcote, 2010; Hawkins, et al., 2014; Little, Nosofsky, & Denton, 2011; Little, Nosofsky, Donkin, & Denton, 2013; Palada, et al., 2016; Trueblood, Brown, & Heathcote, 2014: Vuckovic, Kwantes, & Neal, 2013; Vuckovic, Kwantes, Humphreys, & Neal, 2014). For example, Palada et al. (2016) used LBA accumulators to model a dynamic UAV target classification task in which up to five stimuli appeared on screen at a time, stimuli contained multiple attributes relevant to each decision, and stimuli had asynchronous onset and offset times on the display. Participants performed the task under different levels of time pressure, workload (number of ships on screen), and with varying degrees of dynamic pixel noise affecting the quality of the visual display. Despite the highly dynamic nature of the task and the complexity of the decision, their LBA-based model provided a close fit to data and reproduced the benchmark effects of their experimental manipulations (i.e., lower thresholds under high time pressure, lower drift rates under high pixel noise, lower drift rates for more difficult stimuli). This work suggests that at least in some complex tasks, the assumptions of models of simple choice are approximated well enough to derive good fits and provide psychologically meaningful parameter estimates.

Likewise, Vuckovic, Kwantes, Humphreys, & Neal, (2014) developed an accumulate-to-threshold framework to model how ATC conflict detection performance varied with time pressure (speed vs accuracy instructions) and the difficulty of aircraft stimuli (e.g., by varying speed, distance, approach angle). Their model used specific spatial properties of on-screen aircraft such as airspeed and approach angle to predict the rate of evidence accumulation. The model showed good fits to data and provided parameter estimates with sensible psychological interpretations of threshold and accumulation rate changes consistent with those found in studies of simpler tasks.

Although these studies have had success in modelling complex tasks with dynamic, multi-attribute stimuli, the models often employed complex architectures with highly task-specifc inputs (e.g., spatial properties of on-screen aircraft, perceived distance to crossover point) which do not generalise easily to other applied contexts. In addition, some require highly specialised expiremental designs not easily adapted to applied settings (e.g., Townsend & Nozawa, 1995), and rich parameterisations requiring large amounts of data and computational resources to fit. In addition, none of the aforementioned studies looked at complex tasks involving additional PM demands. As such, there is no computational model of a complex dynamic task that accounts for PM performance in addition to the primary ongoing task.

There have however been several studies that used DM and LBA frameworks to model PM in basic lab paradigms (Boywitt & Rummel, 2012; Heathcote, Loft, & Remington, 2015; Horn & Bayen, 2015; Horn, Bayen, & Smith, 2011; Strickland, Heathcote, Remington, & Loft, 2017, Strickland, et al., 2017). These studies have provided a more detailed analysis of the cognitive processes that underly PM than has been possible with conventrional analsyes of RT and accuracy. However, the models have almost exclusively only been fit non-PM trials rather than the full array of PM data, and as such do not provide a complete account of PM effects (Boywitt & Rummel, 2012; Heathcote, Loft, & Remington, 2015; Horn & Bayen, 2015; Horn, Bayen, & Smith, 2011; Strickland, Heathcote, Remington, & Loft, 2017; see Strickland, et al., 2017 for the an exception).

Rather than attempting to develop more a more complex, task-specific model (e.g., Corker, Gore, Fleming, & Lane, 2000; Eyferth, Niessen, & Spaeth, 2003; Leiden, Kopardekar, & Green, 2003; Neal & Kwantes, 2009; Niessen, Eyferth, & Bierwagon, 1999), our goal here was to evaluate whether a standard choice model with a flexible parameterisation could provide an adequate description of data from a complex and dynmamic task respresentative of those performed in many applied settings.

Conceptual Framework and Modelling Architecture

We used LBA accumulators to model decision-making in a complex dynamic ATC conflict detection task with both time pressure and PM demands. Our model architecture includes three LBA

accumulators: two that correspond to the ongoing task responses (i.e., conflict/nonconflict), and a third that corresponds to the PM response. Figure XX depicts the model for the conflict detection task with a concurrent PM task requirement. There are three possible response alternatives, which correspond to indicating that the stimulus is either a *conflict*, a *nonconflict*, or a *PM target*. Evidence for each response accrues linearly towards threshold, starting from a point that varies independently between accumulators from trial to trial according to a uniform distribution. Correct PM responses (PM hits) occur on PM trials when the PM accumulator reaches threshold before either of the ongoing task (conflict/nonconflict) accumulators. Similarly, PM misses occur when one of the ongoing task accumulators finishes before the PM accumulator. Response probabilities vary depending on the values of three classes of model parameters related to the level of start-point variability and thresholds and evidence accumulation rates. Evidence accumulates at a constant rate within a given trial, but rates differ from trial to trial according to a normal distribution. Accumulation rate parameters are usually assumed to vary as a function of stimulus differences and can vary from trial to trial. Thresholds parameters, in contrast, are set prior to stimulus presentation, and are unaffected by stimulus characteristics that vary unpredictably from trial to trial. However, thresholds can vary over blocked manipulations (e.g., PM vs control blocks; high vs low time pressure), and by response (e.g., a less conservative threshold might be set for 'conflict' responses compared to 'nonconflict' responses). The level of start-point variability is assumed not to vary over the conditions we examine here, although it might vary in other circumstances. Decision time (i.e., the time for the winning response to accumulate to threshold) is determined by the same set of parameters as response probabilities. Total RT is determined by decision time plus nondecision time, which includes all nondecisional processes, such as stimulus encoding and motor response production. Nondecision time was assumed to be the same across all experimental conditions and all response accumulators. Thus we estimated only one nondecision time parameter.

This modelling framework will allow us to test the ability of the LBA to model more complex decisions than it has as yet been applied to, and, assuming good fit, to interpret the underlying cognitive mechanisms driving performance and how they may be affected by changes in time pressure and PM demand. By evaluating model fit, we can determine whether the basic assumptions (i.e., a parallel independent race with feedforward excitation and inhibition and linear updating) are

sufficient to model the processes driving performance in this more complex, dynamic, multi-stimulus applied task. Misfit will suggest that this architecture may not be the correct process model of our more complex task. Misfit will suggest that different strategic (e.g., serial processing, double-checking) or cognitive mechanisms (e.g., nonlinear or piecewise accumulation, collapsing thresholds) may need to be added to the model in order to adequately capture performance. Nevertheless, assuming sufficiently good fit, the way in which model parameters vary to capture the effects of time pressure and PM demand will allow us to evaluate the latent cognitive processes that drive observed responding in this complex, dynamic, multi-stimulus ATC task. The following sections describe the benchmark effects of time pressure and PM demand manipulations, theoretical predictions, and how the proposed model will distinguish between competing theories.

Time Pressure

Human decision-making almost always occurs under some form of time pressure (Svenson & Maule, 1993). Time pressure refers to constraints on the time available to gather information and deliberate before committing to a response or course of action. Time constraints can be internal, in which the decision to stop deliberating and execute a response or action is self-imposed, or external, in which the time available for deliberation is limited by aspects of the task itself (e.g., tasks in which decisions must be made before a deadline) or changes in the task environment (e.g., when increased workload requires more decisions to be made per unit time; Wickelgren, 1977).

Time constraints are extremely prevalent in every-day life and work settings (e.g., deciding on the right moment to merge into traffic; being given a shorter deadline within which to complete a work task), and can affect both the speed and quality with which tasks are performed and decisions made (e.g., having too little time to weigh up all the evidence before committing to a decision; Stokes, Kemper, & Kite, 1997). Increases in time pressure have been implicated in poorer safety outcomes in many safety-critical settings, including road-safety (Gelaua, Sirek, Dahmen-Zimmer, 2011; Shinar, 1998) and aviation (Sarter & Schroeder, 2001).

The effects of time pressure on decision-making are typically studied in experiments that manipulate either the absolute time available for decisions (i.e., by imposing different response deadlines; Frazier

& Yu, 2008; McElree & Dosher, 1989; Meyer Irwin, Osman, & Kounios, 1988) or which vary the subjective importance of fast versus accurate decisions via task instructions or by providing incentives that reward either fast or accurate responding (Milosavljevic, et al., 2010). Importantly, changes in time pressure have systematic effects on RT and accuracy. Higher time pressure (i.e., less deliberation time) generally leads to faster but less accurate decisions, whereas lower time pressure (i.e., more deliberation time) leads to slower but more accurate decisions. This phenomenon is known as the speed-accuracy trade-off (SAT; Liu & Watanabe, 2012; Wickelgren, 1977), and has been the subject of much computational cognitive modelling (Dutilh, Wagenmakers, Visser, & van der Maas, 2011; Forstmann, Tittgemeyer, Wagenmakers, Derrfuss, Imperati, & Brown, 2011; Usher, Olami, & McClelland, 2002) and neuropsychological research (Bogacz, Wagenmakers, Forstmann, & Nieuwenhuis, 2010; Heitz & Schall, 2012, Heitz, 2014).

In terms of formal cognitive theories, the SAT is typically implemented in computational decision models by including a variable response threshold or decision boundary that varies over different levels of time pressure (Ratcliff & Rouder, 1998). By setting a higher (i.e., more conservative) response threshold, operators can choose to gather more evidence or deliberate for a longer period of time before committing to a response, thus producing the characteristic slow-but-accurate response pattern. Conversely, by setting a lower (i.e., less conservative) response threshold, operators spend less time deliberating and gathering evidence, and are therefore able to make faster responses (albeit at the expense of accuracy).

Indeed, allowing the threshold parameter alone to vary across time pressure conditions is often sufficient to allow choice-RT models to closely fit empirical SAT data (e.g., Forstmann, et al., 2008; Ratcliff, 2002; Ratcliff & Rouder, 1998; Ratcliff, Thapar, & McKoon, 2003; Thapar, Ratcliff, & McKoon, 2003; Wagenmakers, Ratcliff, & McKoon, 2008). In contrast, time pressure generally has minimal effect on more automatic decision processes such as accumulation rates and the nondecisional components of response time such as encoding and motor response time (i.e., fixing these parameters has little effect on how well models fit empirical data). Because response thresholds are assumed to be under the conscious control of the decision-maker, these findings suggest time pressure primarily leads operators to make deliberate, proactive adjustments to their response thresholds; responding more conservatively under low time pressure to increase accuracy and less conservatively under

high time pressure to increase speed. In other words, time pressure primarly influences a strategic, proactive component of the decision-making process under the conscious control of the operator.

It should be noted however that time pressure may also influence the quality of evidence accumulation (i.e., accumulation rates; Heathcote & Love, 2012; Heitz & Schall, 2012; Ho, Brown, van Maanen, Forstmann, Wagenmakers, & Serences, 2012; Rae, et al., 2014; Starns, Ratcliff, & McKoon, 2012; Vandekerckhove, Tuerlinckx, & Lee, 2008), nondecision time (Dambacher & Hubner, 2015), and variability in the rate of evidence accumulation (Milosavljevic, et al., 2010). In terms of accumulation rates, the aforementioned studies have suggested that in addition to threshold shifts, time pressure may reduce the quality of evidence entering the decision process (i.e., lower accumulation rates) and increase the variability in accumulation rates reflecting greater noise in the accumulation process (Milosavljevic, et al., 2010). Time pressure may also shorten nondecision time, by speeding up motor response processes (e.g., Dambacher & Hubner, 2015; Osman et al., 2000; Rinkenauer et al., 2004; van der Lubbe et al., 2001), or by increasing temporal preparation, which has been shown to lead to earlier onset times for evidence accumulation processes (Bausenhart, Rolke, Seibold, & Ulrich, 2010; Seibold, Bausenhart, Rolke, & Ulrich, 2011), or by decreasing the frequency of 'double-checking' responses (Boywitt & Rummel, 2012; Guynn, 2003; Horn, Bayen, & Smith, 2011).

In contrast, studies on mind-wandering and distraction suggest that accumulation rates may increase rather than decrease with time pressure (McVay & Kane, 2009; Rummel, Smeekens, & Kane, 2016; Smallwood, 2013; Smallwood & Schooler, 2015). Specifically, more difficult task conditions (e.g., high time pressure) tend to lead to less mind-wandering and increase task focus because participants engage in fewer task-unrelated thoughts while performing the task (Kane & McVay, 2012; Rummel, Smeekens, & Kane, 2016). In the model, task focus maps naturally onto the accumulation rate parameter. As such, if time pressure increases task focus by reducing mind-wandering, we would expect higher accumulation rates in high time pressure blocks compared to low time pressure blocks. Similar predictions are also made by effort mobilisation theories which argue that people have a reserve pool of cognitive resources which they can voluntarily invest into a task by exerting extra effort (Hockey, 1997; Hockey, Coles, & Gaillard, 1986; Kleinsorge, 2001; Sanders, 1983; Schmidt, Kleinbeck, & Brockmann, 1984; Sptiz, 1988; Wickens, 1986). These theories attribute improvements in performance under greater task load to a boost in processing intensity or an increase in signal gain

on evidence sampled from stimuli, rather than a reduction in distracting thoughts (Kleinsorge, 2001; Wickens, 1986). There is strong neurological support for this idea; studies have shown that increased motivation can potentiate neural structures related to cognitive processing (i.e., by increasing the probability of neurotransmitter release), thus facilitating the flow of information (Banquet, Smith, & Guenther, 1992; Beierholm, Guitart-Masip, Economides, Chowdhury, Düzel, Dolan, Dayan, 2013; Botvinick & Braver, 2015; Chiew & Braver, 2013; Gallistel, 1985; Jimura, Locke, & Braver, 2010; Niv, Daw, & Dayan, 2007; Schmidt, Lebreton, Clrey-Melin, Daunizaeu, & Pessiglione, 2012).

In terms of the model, both of these accounts (i.e., decreased distraction and increased effort) would be reflected in the accumulation rate parameter. As such, if processing intensity is increased under high time pressure via effort mobilisation, we would expect higher accumulation rates in high time pressure blocks compared to low time pressure blocks. Considering that our ATC task is more resource-demanding than the basic tasks used in these studies (e.g., lexical decision) people may be already maximally focused, or be expending maximum effort, even during low time pressure blocks. As such, it may be more plausible here for time pressure to negatively affect the quality of information processing (i.e., accumulation rates), since participants may already be committing their full compliment of cognitive resources to the task. Nevetheless, given the novelty of our task, these possibilities are both plausible and will be explored further in our computational modelling.

Time Pressure and Strategic Response Bias

In certain contexts changes in time pressure can also induce strategic shifts in respond bias. Specifically, changes in time pressure may lead operators to begin deliberately prioritising one response over another in order to maintain satisfactory performance on superordinate task goals. Loft, Bolland, Humphreys, and Neal (2009), for example, showed that air-traffic controllers apply larger safety margins by strategically shifting bias towards conflict responses when placed under greater time pressure in order to satisfy the superordinate task goal of maintaining the safe passage of aircraft (Loft, et al., 2009). Although this strategy increases the rate of conflict false alarms, controllers adopt it because the goal of ensuring aircraft safety is more important than strictly accurate identification of conflicts versus nonconflicts (Loft, et al., 2009). In terms of the model, response bias is reflected in differences in response thresholds between competing ongoing task task

responses. Lowering the thresholds for one response relative to another means the first response requires less evidence to trigger a response, resulting in a bias towards that response. In the present task we expect higher time pressure will lead participants to prioritise conflict responses over nonconflict responses which will be reflected in conflict thresholds being lowered relative to nonconflict thresholds as time pressure increases (Loft, et al., 2009).

Prospective Memory Demands

In addition to time constraints, much real-world decision-making is performed under prospective memory demand (Dismukes, 2012). Prospective memory refers to the formation and maintenance of intentions to perform future actions - actions which cannot be completed immeditately and must be deferred either until a later point in time (i.e., *time-based PM*; Park, et al., 1997) or until a particular future event or stimulus is encountered (i.e., *event-based PM*; Brandimonte, Einstein, & McDaniel, 1996; Einstein & McDaniel, 1990; Ellis & Cohen, 2008; Kliegel, McDaniel, & Einstein, 2008). The period between encoding and executing PM intentions is usually filled with one or more ongoing tasks unrelated to the intention, which typically must be retrieved and executed in the absence of an explicit reminder (Dismukes, 2012).

Event-based PM task requirements are highly prevalent in everyday life (e.g., remembering to send an email once you arrive at work, remembering to take food out of the oven), and can have important satefy implications (e.g., remembering to clear one aicraft off the runway before allowing another to land; National Transportation Safety Board, 1991; remembering to slow down while driving through a school zone; Bowden, Visser, & Loft, 2017). In aviation research, for example, addtional PM demand has been implicated in operators being slower to accept and handoff aircraft, slower to detect conflicts between aircraft, as well as increased rates of missed conflicts (Loft, Finnerty & Remington, 2011; Loft, Pearcy, & Remington, 2011; Loft & Remington, 2010; Loft, Smith, & Bhaskara, 2011; Loft, Smith, & Remington, 2013; Loukopoulos, Dismukes, & Barshi, 2009). Reliable retrieval and execution of PM intentions is therefore crucial to performance in safety-critical decision-making contexts such as aviation, medicine, and defence (Dembitzer & Lai, 2003; Dismukes, 2012; Gawande, Studdert, Orav, Brennan, & Zinner, 2003; Grundgeiger & Sanderson, 2009; Loft, 2014). Importantly, because operators may perform many thousands of actions under PM load per day, even small

PM error probabilities have the potential to translate into significant accident rates (Dismukes & Nowinski, 2006; Shorrock, 2005). As such, it is critical that models of applied tasks properly account for the effects of concurrent PM task demand on operator performance.

The effect of PM demand on decision-making is typically studied in experimental paradigms in which an infrequently-occurring PM task is embedded within a primary 'ongoing' task (Einstein & McDaniel, 1990). For example, a common ongoing task used to study PM is the lexical decision task, in which participants must decide whether letter strings are words or not. Prior to performing the task, participants are instructed to make an atypical PM response (e.g., press 'F1') if they encounter a PM target (e.g., any words related to the semantic category 'animal'). Participants typically complete both a control block in which the ongoing task is performed by itself with no PM requirement, and a PM block in which the ongoing task is performed concurrently with the embedded PM task. Ongoing task performance can then be compared with and without PM demand (Smith, Hunt, McVay, & McConnell, 2007).

A robust finding in the PM literature is that ongoing task RTs are typically longer in PM blocks compared to control blocks. That is, people make slower ongoing task responses when required to hold concurrent PM intentions, a phenomenon known as *PM cost* (Einstein et al., 2005; Hicks, Marsh, & Cook, 2005; Loft & Yeo, 2007; Smith, 2003).

PM costs have been used to infer the cognitive processes that underlie PM performance as well as those believed to drive observed ongoing task costs. For example, capacity-sharing theories of PM costs (Craik, 1986; Marsh & Hicks, 1998; Park, Hertzog, Kidder, Morrel, & Mayhorn, 1997) attribute PM costs to some limited-capacity cognitive resource (e.g., prepatory attention, working memory) being shared between the ongoing and PM tasks (Boywitt & Rummel, 2012; Einstein & McDaniel, 2005; Smith, 2003). That is, the processes responsible for holding PM intentions or monitoring for PM targets are assumed to draw cognitive resources away from the ongoing task, thereby reducing the efficiency with which the ongoing task can be performed (Smith, Hunt, McVay, & McConnell, 2007). In terms of evidence accumulation models, processing efficiency maps naturally on to the drift rate or evidence accumulation rate parameters. Computationally then, the capacity-sharing account attributes PM costs to a bottom-up, stimulus-driven reduction in ongoing task accumulation rates (i.e., less efficient processing) under PM load.

Standard capacity-sharing theories and the MPV have recently been challenged, however, by recent work showing that PM costs are primarly the result of strategic, metacognitive adjustments to the task environment rather than cognitive processing limitations (Heathcote, Loft, & Remington, 2015; Strickland, 2017). This theory (referred to here as *delay* theory) argues that under PM load people set more conservative ongoing task response thresholds in order to avoid preepmting rare PM targets with the more habitual ongoing task responses. Setting more conservative ongoing task thresholds results in longer ongoing task RTs under PM load (i.e., PM costs), which facilitates responses to PM stimuli while maintaining comparable ongoing task accuracy between control and PM blocks (Strickland, 2017). That is, people may make similar strategic, proactive adaptations in response to PM demands as they do to different levels of time pressure. Other accumulation model parameters such as nondecision time are not typically implicated in PM processing (but see Boywitt & Rummel, 2012 who have argued that additional checking and/or encoding of PM stimuli may be reflected in a longer nondecision component; also see Guynn, 2003; Horn, Bayen, & Smith, 2011).

Proactive and Reactive Cognitive Control Mechanisms

In addition to top-down, deliberate control of thresholds in response to PM demand (referred to as *proactive control*), Strickland et al. (2017) found evidence of bottom-up or stimulus-driven inhibitory processes acting between PM and ongoing task stimuli. These processes are known as *reactive control* and operate automatically rather than being under the decision-maker's conscious or deliberate control (Braver, 2012; Bugg, McDaniel, & Einstein, 2013; Ball, 2015). Specifically, Strickland et al. (2017) found that evidence accumulation rates for ongoing task responses were significantly lower on trials that also contained PM stimuli than on trials that did not contain PM stimuli, despite both trial types having equally strong evidence for the respective ongoing task response. They argued that although both PM and non-PM trials contained equal evidence for the ongoing task, on PM trials the ongoing task accumulators received inhibitory input from the PM detector, which reduced their rates of accumulation (Strickland et al., 2017). Because the presentation of PM stimuli was random, meaning participants could not make deliberate anticipatory processing adjustments (e.g., to prepatory attention), reactive inhibition appears to be a largely automatic mechanism of decision control.

This finding, along with the benchmark effect of proactive threshold control, led Strickland et al., (2017) to frame their modelling results in terms of Braver's (2012) dual-mechanisms theory of cognitive control, which argues that human decision-making is subject to both proactive and reactive control mechanisms. Braver's (2012) theory has been useful for understanding cognitive control in many paradigms, including working memory tasks (e.g., Braver, Gray, & Burgess, 2007; Burgess & Braver, 2010; Marklund & Persson, 2012), the AX-Continuous performance task (e.g., Braver, Barch, Keys, et al., 2001; Locke & Braver, 2008; van Wouwe, Band, & Ridderinkhof, 2011), the stop signal task (e.g., Stuphorn & Emeric, 2012; Boehler, Schevernels, Hopf, Stoppel, & Krebs, 2014), the Stroop task (e.g., Kalanthroff, Avnit, Henik, Davelaar, & Usher, 2015; West, Choi, & Travers, 2010), and the cued task-switching paradigm (e.g., Chevalier, Martis, Curran, & Munakata, 2015; Lucenet, Blaye, Chevalier, & Kray, 2014). Additional empirical support for the dual-mechanisms account comes from psychophysiological studies showing that proactive and reactive control mechanisms correspond to distinct patterns of brain activity (Braver, 2012; Irlbacher, Kraft, Kehrer, & Brandt, 2014; Appelbaum, Boehler, Davis, Won, & Woldorff, 2014). The following section briefly outlines proactive and reactive control mechanisms in terms of Braver's (2012) dual-mechanisms framework and their relevance to the present study.

Proactive Control

In Braver's (2012) framework, proactive control refers to processes used to "bias attention, perception and action systems in a goal-driven manner" (Braver, 2012, p. 2). Proactive processes are deployed deliberately, in an anticipatory manner. That is, they are activated in advance of the target stimulus or event so that they will already be active when the target stimulus is encountered (Braver, 2012). Because they are assumed to be under conscious control, proactive control processes are likely to be deployed for entire blocks of trials rather than on a trial-to-trial basis (especially when trials are presented randomly and cannot be anticipated).

In terms of time pressure and PM demand, proactive control in response to these manipulations would be active on all trials within a time pressure or PM demand block (e.g., all high time pressure trials, all PM block trials). Importantly, because proactive control processes selectively map onto the threshold parameter in accumulate-to-threshold models, their effects can be separated out from

other non-proactive processes that may have similar block-wise effects on RT and accuracy (e.g., capacity cost effects). For example, longer RTs under PM load (i.e., PM costs) are predicted by both capacity-sharing and strategic delay theories of PM cost. Although both theoretical processes may have similar effects on manifest RT, the proposed source of those effects is very different. Capacity theories attribute PM block slowing to poorer-quality information processing due to resources being diverted from the ongoing task - which in the model would be reflected in the rate of evidence accumulation. In contrast, delay theories attribute slowing to more cautious responding - the result of setting higher ongoing task thresholds to give the parallel PM accumulation process a better chance of reaching response selection (Heathcote et al., 2015; Loft & Remington, 2010). Although these two accounts are confounded in conventional analyses of mean RT, they can be identified in the present model-based analysis.

Reactive Control

In contrast to deliberate, proactive control mechanisms deployed in anticipation of task demands, Braver (2012) argues that there are also automatic, stimulus-driven cognitive mechanisms that are deployed to influence responding "only as needed, in a just-in-time manner" (Braver, 2012, p. 2). These are referred to as reactive control mechanisms. Reactive control mechanisms are 'bottom-up', automatic cognitive processes and are assumed to be largely outside of conscious control. As such, reactive control processes can be deployed on a trial-to-trial (or stimulus-to-stimulus) basis; they are not restricted to entire blocks of trials.

In terms of PM tasks, reactive control processes would be activated on PM trials when a PM stimulus is present but remain inactive on non-PM trials when no PM stimuli are present. An example of a reactive control process that would facilitate PM responding is if ongoing task processing is inhibited when PM stimuli are present. This kind of inhibition would improve PM accuracy. Due to their automatic, stimulus-driven nature, reactive control processes map selectively onto the accumulation rate parameter in sequentially sampling frameworks. Comparing accumulation rate parameters between PM trials and non-PM trials (in PM blocks only) allows us to detect and quantify reactive control effects. Moreover, because reactive control predicts different rates for PM and non-PM trials, it can be distinguished from capacity sharing, which predicts lower accumulation rates on all PM

block trials regardless of whether a PM stimulus was presented or not. The next section describes in detail how we propose to use our model to identify and test these effects in a complex, dynamic ATC task.

The Current Experiment

This study tests the ability of our model (based on the PMDC architecture of Strickland et al., 2017) to provide full account of decision-making under time pressure and PM demand in a complex, dynamic air-traffic control task. To this end, we have several objectives. First, we evaluate whether the model is capable of fitting the full array of data across our time pressure and PM demand manipulations. Adequacy of fit will indicate whether the LBA assumption that PM and ongoing task decision processes compete in an independent race is reasonable in this complex applied task. Model misfit could also indicate that the LBA is the wrong model of each accumulation process. Indeed, this is a plausible scenario in applied tasks as complex and dynamic as ours, in which decision-making processes may operate differently than in basic cognitive and perceptual lab tasks. For example, because our ongoing conflict detection task is dynamic, meaning that stimulus evidence unfolds over time, the task may violate the core LBA assumption that evidence accumulates linearly within each trial. Nevertheless, given the breadth of successful applications of the LBA in a wide variety of basic and applied paradigms (e.g., Brown & Heathcote, 2008; Eidels, Donkin, Brown, & Heathcote, 2010; Forstmann et al., 2008; Matzke, Love, & Heathcote, 2017; Palada, Neal, Vuckovic, Martin, Samuels, & Heathcote, 2016; Provost & Heathcote, 2015), and recent work suggesting constant-rate models are good approximations even if accumulation rates do change within trials (Ratcliff, 2002; Ratcliff, Smith, Brown, & McKoon, 2016), this seems like a relatively unlikely possibility. Second, assuming we find good fit, we will interpret model parameters in order to evaluate how well the data are described by capacity sharing theories of PM cost versus PMDC and strategic delay theory accounts. This can be done in part by comparing control blocks to PM blocks (proactive control), and in part by comparing non-PM trials with PM trials (reactive control). Moreover, we will evaluate how thresholds and accumulation rates change with time pressure, and whether proactive control of thresholds interacts with time pressure in any way.

Our primary goal when designing this experiment was to reliably model both ongoing task and PM

responses in a more complex and dynamic task than is typically used in either choice-RT modelling or PM research. To this end, our task deviates from traditional PM paradigms in several ways. First, most PM studies conducted using the traditional Einstein and McDaniel (1990) paradigm present only a small number of PM stimuli across an experiment, and as such they do not produce enough data to reliably constrain a model of PM processes. To reliably constrain our model, and increase the power of model fitting, we modified the paradigm by increasing the ratio of PM target trials to ongoing task trials to 1 in 5 (i.e., 20% of PM block stimuli were PM targets). This is a higher proportion of PM target trials than is typically used in PM literature, however, given that the present task is much more complex than traditional PM tasks (such as lexical decision), that decisions in this task unfolded over a much longer timescale (i.e., up to 10 seconds), and that overall PM accuracy was comparable to that of previous PM literature (~72%), the higher PM target ratio appears to be a reasonable modification to the task.

Second, we instructed participants to make their PM response instead of their ongoing task response, as we have done in our previous computational modelling efforts (Heathcote et al. 2015; Strickland et al., 2017). We used this type of PM because it is relevant to everyday situations in which cognitive errors are a result of failing to inhibit routine actions (Norman, 1981; Reason, 1990). In addition, this means we record only one RT and one response on every trial, allowing us to fit the standard LBA, whereas fitting data from a paradigm where both PM and ongoing task responses can be submitted to the same stimulus would require a more complex modelling approach.

Third, we modified the response key arrangements from the typical paradigm. Usually in the PM literature participants rest their fingers on the ongoing task keys and have to make a larger movement for the PM response. In terms of the LBA, this would cause uneven nondecision time between responses (via motor response production time), which adds complexity to the model that is not relevant to understanding PM processes, and neglecting unequal nondecision times between responses can lead to biased estimates of other model parameters (Voss, Voss, & Klauer, 2010). As such, we instructed participants to rest their fingers on both ongoing task response keys (with one hand), and on the PM response key (with the other hand), so we could assume an equal motor response time.

To assess the effects of time pressure and PM load on task performance, we used a repeated measures

design in which participants completed a simulated ATC conflict detection task under two PM load conditions: control (no additional PM task) and PM (additional requirement to respond to PM targets), and four time pressure/trial load conditions: low load/low time pressure, low load/high time pressure, high load/low time pressure, and high load/high time pressure. Because we model all response types on both PM and non-PM trials, this design allows us to examine how ongoing task and PM response thresholds (i.e., proactive control) and accumulation rates (i.e., reactive control) might change under different time pressure and PM load conditions.

Our approach to modelling and analysis deviates from traditional methods used in the PM literature in several ways. First, we estimate parameters using Bayesian methods, whereas most previous modelling (the exception being Strickland, 2017) has used optimisation methods like maximum likelihood. This allows us to obtain full probability distributions of likely values for each model parameter, rather than single point estimates. Second, in addition to model selection, we used posterior inference, in which conclusions are drawn based on comparisons of parameters posterior distributions (see Brooks & Gelman, 1998). The following sections outline how the various theoretical accounts predict specific model parameters will behave under time pressure and PM demand, and how we will determine whether evidence for SAT effects, capacity-sharing, proactive, and reactive control is present in our data.

Selective Influence of Time Pressure

Following the large literature on the speed-accuracy trade-off, we expect increased time pressure to be primarly reflected in decreased thresholds for both ongoing task and PM responses. That is, when there is less time available to make decisions, thresholds will be lowered in order to facilitate faster responding. In terms of the effects of time pressure on other model parameters, particularly accumulation rates, the evidence is mixed. Assuming time pressure decreases the quality of information processing, we would expect lower accumulation rates under high time pressure (Ho, Brown, van Maanen, Forstmann, Wagenmakers, & Serences, 2012). In contrast, assuming time pressure leads participants to become more less distracted/more focused on the task or to invest greater effort, we would expect higher accumulation rates in high time pressure blocks compared to low time pressure blocks (Dambacher, Hubner, & Schlosser, 2011; Dambacher & Hubner, 2013;

Hubner & Schlosser, 2010; Kleinsorge, 2001). These two effects are not mutually exclusive: one might occur, neither might occur, both might occur and cancel each other out, or both might occur with one having a stronger effect than the other. As such, we cannot unequivocably rule either out. Nevertheless, lower rates under time pressure would suggest a time-pressure related cost to processing efficiency regardless of increased effort, while higher rates under time pressure would suggest an increase in effort or arousal, regardless of processing costs.

In terms of time pressure, our design is similar to research on the speed-accuracy trade-off in which participants can choose to make faster responses at the expense of accuracy or make more accurate responses at the expense of speed. This research typically manipulates time pressure by imposing response deadlines of different durations or through task instructions that emphasise the importance of either fast or accurate responding. A benchmark finding in this literature is that higher time pressure (i.e., speed emphasis) leads people to lower their response thresholds to facilitate faster responding. Conversely, low time pressure (or accuracy emphasis) leads participants to raise their response thresholds in order to gain accuracy. We expect to replicate these findings for both ongoing task and PM thresholds. In addition, we expect higher time pressure to lead to a shift in response bias in which conflict responses are prioritised over nonconflict responses. Specifically, we expect this bias to manifest in a deliberate reduction in conflict thresholds relative to nonconflict thresholds.

Selective Influence of PM Demands

In terms of PM load, our design is similar to Strickland (2017) and Heathcote, Loft and Remington (2015), which modelled responding to non-PM trials and found that PM costs were the result of strategic increases in response thresholds under PM load rather than decreased accumulation rates predicted by capacity-sharing theories of PM cost. Consistent with these results we expect thresholds for both ongoing task (conflict/nonconflict) responses to be higher during PM blocks than in control blocks. Concerning accumulation rates under PM load, Strickland (2017) found that ongoing task accumulation rates were lower on PM target trials compared to non-PM trials, suggesting the PM detector inhibits competing ongoing task responses via an automatic, stimulus-driven, 'reactive' control process. We expect to replicate these findings, although given the more complex and dynamic nature of our task and that we used a completely non-focal PM cue (non-focal PM cues generate

less reactive inhibition), it is not immediately clear whether these effects will replicate as strongly.

Strategic Response Bias

In line with Loft et al. (2009), we expect an increased bias towards conflict responses under high time pressure compared to low time pressure conditions. We expect this bias to be reflected in a reduction in conflict thresholds relative to nonconflict thresholds in high time pressure blocks. A shift in response bias toward conflicts under high time pressure would suggest time pressure leads participants to prioritise the superordinate task goal of ensuring aircraft safety over maintaining strict accuracy in terms of discriminating conflicts from nonconflicts.

Capacity-Sharing

PM capacity demand on the ongoing task would be reflected in decreased non-PM trial accumulation rates in PM blocks. As reviewed, this account has been rejected for at least eight previous data sets (Heathcote et al., 2015; Horn & Bayen, 2015; Strickland et al., 2017). However, given the novelty of our more complex and dynamic ATC task, and the fact that we are using a completely non-focal PM cue which would have a higher capacity requirement according to capacity-sharing theories (Einstein & McDaniel, 2005; Smith, Hunt, McVay, & McConnell, 2007), it is possible that we may find evidence of capacity sharing.

However, another possibility raised in recent work on mind-wandering in PM tasks (Rummel, Smeekens, & Kane, 2016; Kane & McVay, 2012; McVay & Kane, 2009; Smallwood, 2013; Smallwood & Schooler, 2015) is that the additional demands of the PM task reduce the frequency of task-unrelated thoughts, thereby increasing task focus during PM blocks. Assuming engaging in thoughts unrelated to the task at hand negatively affects the efficiency of information processing, we would expect increased ongoing task accumulation rates in on non-PM trials in PM blocks (in which less mind-wandering occurs) and lower ongoing task accumulation rates in control blocks (in which more mind-wandering occurs; Kane & McVay, 2012) - essentially the opposite effects to what would be predicted under capacity demand. As with the accumulation rate predictions related to time pressure, these possibilities are also not mutually exclusive. Lower rates under PM load would

suggest capacity demands but not rule out increased effort or task focus, while higher rates under PM load would suggest increased effort or focus but not rule out capacity demands.

Proactive Control

Strategic delay and PMDC theories predict ongoing task thresholds will be proactively controlled in PM blocks in order to facilitate PM responding. In line with previous findings (Heathcote et al., 2015; Strickland et al., 2017), we expect that ongoing task thresholds will be higher in PM blocks compared to control blocks. Our design allows both ongoing task stimuli (conflicts and nonconflicts) to be PM targets, and PM cues are divided evenly between the two stimulus types. This is expected to lead to similar competition between the PM response and both ongoing task responses, and as such we expect to see similar threshold increases for both response types (conflict and nonconflict).

Reactive Control

A further prediction of PMDC is that that ongoing task accumulation rates can be reactively controlled (inhibited) on PM trials as compared with non-PM trials (Strickland, et al., 2017). Naturally, the PM accumulation rate should be higher on PM trials due to the processing of PM related stimulus attributes that are not present on non-PM trials. In addition however, rates for the ongoing task accumulators are expected to be lower on PM trials. This is because although both PM and non-PM trials contain equal evidence that a stimulus is a conflict or nonconflict, on PM trials the accumulators for the ongoing task responses may receive additional inhibitory input from the PM detector, thus reducing their rates of accumulation (Strickland, et al., 2017). As such, we expect ongoing task rates to be lower on PM trials, which will provide positive evidence that reactive control processes are present.

Method

Participants

Two participants were excluded (see Results), with 47 participants remaining (31 females). Ages ranged from 18 to 62 years ($M = 25.19$, $SD = 9.99$). Participants completed one two-hour testing session.

Materials

ATC Conflict Detection Task

The ATC conflict detection task (Fothergill, Loft & Neal, 2009) involved classifying pairs of moving aircraft as either 'conflict' or 'nonconflict' depending on whether the aircraft would violate a 5 nautical mile (NM) minimum separation distance at some point during their flight. On some trials aircraft also contained a PM cue which required execution of a PM response instead of a conflict or nonconflict ongoing task response.

During the simulation, the visual display included a countdown timer indicating seconds remaining in the trial and a 10NM by 20NM (approximately 2cm by 4cm on screen) scale marker to be used as a reference when judging relative aircraft distances and separation. An information block showing callsign, airspeed, and flight level (altitude) information tracked alongside each aircraft, which appeared on screen as small circles. Each aircraft had a probe vector line which showed the aircraft's heading and predicted position one minute into the future. All aircraft appeared within a circular ATC sector with a neutral grey background and followed converging straight-line flight paths (indicated by black lines) which always crossed over at a 90-degree angle in the center of the display. We now describe the spatial parameters of the aircraft stimuli used in the ATC simulation, the PM cue, and experimental design.

Aircraft Stimuli

We generated a set of unique conflict and nonconflict aircraft stimuli for each participant. To create the different conflict and nonconflict stimuli, each aircraft pair was assigned a distance of minimum separation (DOMS) either less than or greater than the 5NM separation standard. DOMS for conflict stimuli were drawn from the uniform distrubtion [0,3] NM. DOMS for nonconflict stimuli were drawn from the uniform distribution [7,10] NM.

In addition to DOMS, we allowed several other spatial properties of each aircraft to vary randomly during stimulus generation. Specifically, airspeed, direction of approach, time to minimum separation (TTMS), and order of passing (i.e., faster aircraft first vs. slower aircraft first) all varied randomly between stimuli. This was done to avoid instance learning of the conflict detection task (see Loft, Humphreys & Neal, 2004; Bowden & Loft, 2016), which can cause problems for modelling by introducing non-stationarity in model parameters. The angle of approach between aircraft was fixed at 90 degrees to avoid interactions between angle and perceived conflict status (Loft, Bolland, Humphreys & Neal, 2009). Finally, the flight level for all aircraft was fixed at 37,000 feet. Table 1 specifies the range of values for each spatial variable and the distribution they were drawn from to generate the stimuli.

Table 1: Range and Distribution of Spatial Variables of Aircraft Stimuli

Spatial Variable	Distribution	Lower	Upper	Units
DOMS (Conflicts)	Uniform	0	3	NM
DOMS (Nonconflicts)	Uniform	7	10	NM
Airspeed	Uniform	400	700	Mph
Angle of approach	Constant	90	90	Degrees
Direction of approach	Uniform	0	360	Degrees
Flight level	Constant	37,000	37,000	Feet
TTMS	Uniform	120	210	Seconds

Spatial Variable	Distribution	Lower	Upper	Units
Order of passing	Bernoulli	0	1	0 = fastest first, 1 = slowest first

PM Cue

Aircraft with callsigns containing two of the same letter (e.g., APA169, RTR451) were designated as PM targets. This is an ecologically valid PM cue, since air-traffic controllers may be asked to look out for a specific flight number and perform an intended action (e.g., put the aircraft in a holding pattern) when that cue occurs. We note that this PM cue is completely non-focal, meaning the evidence used to make PM decisions (i.e., particular letters in an aircraft callsign) is independent of evidence used to make ongoing task conflict/nonconflict decisions (e.g., airspeed, relative distance, and motion information). In our experiment, 20% of PM block trials contained a PM target. On PM target trials only one of the aircraft on screen ever contained a PM cue, never both. Participants were instructed to respond to PM targets by pressing an alternate PM key instead of the typical ongoing task (i.e., conflict/nonconflict) keys.

NASA-TLX

Subjective task demand was assessed using the NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988). The NASA-TLX comprises six items: three that tap different aspects of task demand (i.e., mental, physical, and temporal demand) and three which tap the operator's subjective level of exertion and task performance (i.e., performance, effort, and frustration). Each item is rated on a 21-point numerical scale, with higher scores being indicative of a more demanding or effortful task. The NASA-TLX is a sensitive, reliable, and valid measure of workload (Hart & Staveland, 1988; Xiao, et al., 2005), and has been used in a wide variety of basic and applied tasks (Hart, 2006).

Experimental Design

Participants performed 8 blocks of trials consisting of four 80-trial control blocks paired with four 240-trial PM blocks, completed in alternating sequence. The order in which blocks were presented (i.e., control or PM first) was counterbalanced across participants.

In control blocks, participants were presented with a randomised sequence of 80 aircraft pairs (40 conflict and 40 nonconflict), with no PM stimuli. In PM blocks, participants were presented with a randomised sequence of 240 aircraft pairs (120 conflict and 120 nonconflict). Of these, a random 48 (24 conflict and 24 nonconflict) also contained a PM cue. Thus 20% (48/240) of PM block stimuli were PM targets. This eight-cell design is shown in Table XX.

Table 2: Details of Experimental Blocks with Number of Control and PM Aircraft Presented

		Control Block	
Trial Load	Time Pressure	Aircraft	PM Block Aircraft
Low	Low	80	240
Low	High	80	240
High	Low	80	240
High	High	80	240

Our time pressure and trial load factors were also blocked. Specifically, each Control-PM block pair had an associated trial load of either 2 or 5 decisions per trial (i.e., 2 or 5 aircraft pairs presented per trial) crossed with an associated level of time pressure (i.e., low or high - corresponding to either a long or short response deadline). This resulted in 4 unique trial load by time pressure combinations, with presentation order counterbalanced across participants. Table 2 shows the details of our time pressure by trial load manipulation.

We note that time pressure was not crossed orthogonally with trial load. That is, under low trial load (2 decisions per trial), low time pressure corresponded to a response deadline of 12 seconds (i.e., 6 seconds per decision on average) while high time pressure corresponded to a response deadline of

8 seconds (i.e., 4 seconds per decision on average). In contrast, under high trial load (5 decisions per trial), low time pressure corresponded to a response deadline of 20 seconds (i.e., 4 seconds per decision on average) while high time pressure corresponded to a response deadline of 10 seconds (i.e., 2 seconds per decision on average).

Table 3: Details of Trial Load and Time Pressure Manipulation

Block	Trial Load	Time Pressure	Decisions per Trial	Response Deadline (s)	Seconds per Decision
1	Low	Low	2	12	6
2	Low	High	2	8	4
3	High	Low	5	20	4
4	High	High	5	10	2

Procedure

Each testing session consisted of a training phase and an experimental phase which altogether took approximately 2 hours to complete. During the training phase participants received verbal instructions about the ATC task, watched an on-screen demonstration of the task environment, and completed a block of 40 training trials which included corrective feedback after each response. During the experimental phase participants completed eight blocks of experimental trials which did not include feedback.

For the ATC task, participants' primary task was to judge whether pairs of aircraft would violate the 5 nautical mile (NM) minimum separation standard at any point during their flight. Participants were instructed that aircraft passing within the 5NM distance were defined as conflicts and required a *conflict* key-press response, whereas aircraft passing outside of that distance were defined as nonconflicts and required a *nonconflict* key-press response. Participants were told that each aircraft pair would be presented sequentially (i.e., only two aircraft would appear on screen at a time), that all aircraft would be moving towards each other on converging flight paths which crossover in the center of the display, and that a number of spatial properties of the aircraft would vary from trial to

trial, including their relative airspeeds, distance of minimum separation, and starting distance from the central crossing point. Participants were instructed to consider these variables when forming their decisions and to refer to the on-screen distance scale and airspeed displays in order to make the best judgment possible. Before each block of trials, participants saw visual instructions reminding them of the trial time limit (time pressure) and the number of aircraft pairs to be presented in each trial (trial load). Depending on the PM block, participants received either control or PM instructions. Before control blocks, participants were instructed that they only needed to make conflict/nonconflict responses for that block. Before PM blocks, participants were instructed to press a PM repsonse key instead of the conflict or nonconflict keys when they saw a PM target. Participants completed a short distractor task and saw a final reminder to respond as quickly and accurately as possible, after which the primary ATC task trials began.

Four response key assignments were counterbalanced across participants; 1) s = conflict, d = nonconflict, j = PM, 2) d = conflict, s = nonconflict, j = PM, 3) k = conflict, j = nonconflict, d = PM, and 4) j = conflict, k = nonconflict, d = PM. Participants were instructed to rest their fingers on their particular response key combination throughout the task. A screen with the text 'Press [Space] to continue' preceeded each trial, and each trial began once the spacebar was pressed. Within each trial, pairs of aircraft were presented sequentially; each pair disappearing from the screen once a response was made.

Trials ended when either all aircraft pairs had been responded to (2 pairs during low-load trials; 5 pairs during high-load trials) or when the response deadline expired (i.e., the timer counted down to zero). Aircraft pairs not responded to within the response deadline were recorded as nonresponses. Aside from the training trials, no further feedback was given concerning task performance. Participants completed a NASA-TLX workload questionnaire after each block. Participants took self-paced breaks between each block of trials and were also permitted short breaks at any point between-trials if required.