Journal of Experimental Psychology: Applied
2016, Vol. 22, No. 1, 1–23

1076-898X/16/$12.00 http://dx.doi.org/10.1037/xap0000074

Evidence Accumulation in a Complex Task: Making Choices About Concurrent Multiattribute Stimuli Under Time Pressure

Hector Palada and Andrew Neal
The University of Queensland

Anita Vuckovic
Advanced VTOL Technologies, Melbourne, Victoria, Australia

Russell Martin
Defence Science and Technology Organization, Melbourne, Victoria, Australia

Kate Samuels
The University of Queensland

Andrew Heathcote
The Universities of Tasmania and Newcastle

Evidence accumulation models transform observed choices and associated response times into psychologically meaningful constructs such as the strength of evidence and the degree of caution. Standard versions of these models were developed for rapid (~1 s) choices about simple stimuli, and have recently been elaborated to some degree to address more complex stimuli and response methods. However, these elaborations can be difficult to use with designs and measurements typically encountered in complex applied settings. We test the applicability of 2 standard accumulation models—the diffusion (Ratcliff & McKoon, 2008) and the linear ballistic accumulation (LBA) (Brown & Heathcote, 2008)—to data from a task representative of many applied situations: the detection of heterogeneous multiattribute targets in a simulated unmanned aerial vehicle (UAV) operator task. Despite responses taking more than 2 s and complications added by realistic features, such as a complex target classification rule, interruptions from a simultaneous UAV navigation task, and time pressured choices about several concurrently present potential targets, these models performed well descriptively. They also provided a coherent psychological explanation of the effects of decision uncertainty and workload manipulations. Our results support the wider application of standard evidence accumulation models to applied decision-making settings.

Keywords: response time, linear ballistic accumulator model, diffusion model, workload, decision uncertainty

Over the last 50 years, detailed and psychologically coherent accounts of rapid choice have been developed based on a simple idea: Decision-makers select a response when they accumulate a threshold amount of evidence favoring it. In a wide range of paradigms where choices involve relatively simple classification or target detection (e.g., Is the stimulus moving left or right? Is it a word? Was it recently encountered?), with responses that are relatively rapid (typically less than 1 s), models based on this idea have provided a fine-grained description of decision-maker's behavior (e.g., Rae, Heathcote, Donkin, Averell, & Brown, 2014). This description encompasses not only the probability of making a choice but also the full distribution of choice response times.

However, it is unclear whether accumulate-to-threshold models provide an accurate description of performance on the types of complex decision tasks found in applied settings. There are certainly similarities; surveillance tasks, for example, often require the operator to detect targets. However, rather than a single simple stimulus property, complex combinations of attributes often define targets, so that even after practice they can take several seconds to discriminate from nontargets. Instead of displays with only one potential target, several can be simultaneously present. Stimuli may only be available for a limited time, appearing and disappearing asynchronously, and other tasks can interrupt, so decision-makers must continually prioritize their information processing and responding. We ask whether two standard accumulate-to-threshold models, the diffusion model (Ratcliff & McKoon, 2008) and the linear ballistic accumulation (LBA) model (Brown & Heathcote, 2008), can approximate the fine-grained details of behavior in such complex decision-making tasks.

This article was published Online First February 4, 2016.

Hector Palada and Andrew Neal, School of Psychology, The University of Queensland; Anita Vuckovic, Human Factors, Advanced VTOL Technologies, Melbourne, Victoria, Australia; Russell Martin, Aerospace Human Factors, Defence Science and Technology Organization, Melbourne, Victoria, Australia; Kate Samuels, School of Psychology, The University of Queensland; Andrew Heathcote, Schools of Medicine and Psychology, The Universities of Tasmania and Newcastle.

Australian Research Council Discovery Project DP120102907 and Professorial Fellowship DP110100234 supported Heathcote. We would like to thank Daniel White for his programming and technical support of the UAV task.

Correspondence concerning this article should be addressed to Hector Palada, School of Psychology, The University of Queensland, St Lucia QLD 4072, Australia. E-mail: hector.palada@uqconnect.edu.au

Accumulate-to-threshold models can be seen as dynamic measurement models that, like signal detection theory signal detection theory (SDT; Green & Swets, 1966), can account for response biases. However, they provide a more precise interpretation of performance by taking account of temporal information. Temporal information allows them to account for complex interactions between response speed and accuracy in terms of latent psychological processes. First, the models quantify the evidence accumulation process, which indexes the quality of the information available to the decision-maker and the degree to which they attend to that information. Second, estimates of the decision threshold index strategic variations in the amount of evidence required by the decision-maker to trigger response production. Finally, observed responses times are divided into a period for response selection and a period for nondecision processes (e.g., stimulus encoding and response production). This refinement in interpretation has important applied implications, such as determining whether errors are produced by the quality of information provided by an interface or the operators' strategies in using it.

Accumulate-to-threshold models have also been developed for specific applied tasks. Neal and Kwantes (2009) developed a model for air-traffic control (ATC) conflict detection, and found that the model was able to account for the complex effects of conflict geometry and timing on accuracy and mean response times (see also Vuckovic, Kwantes, & Neal, 2013). In a follow-up study, they found that the criterion and threshold parameters responded in a theoretically consistent way to manipulations of bias and speed-accuracy, respectively (Vuckovic, Kwantes, Humphreys, & Neal, 2014). Although this work reveals the applicability of accumulate-to-threshold models to complex tasks, it is of limited generality as the model is specific to conflict detection in ATC. Inputs to the model include task specific parameters such as aircraft velocity, perceived distance of aircraft crossing points, and controller's safety margins. In contrast, standard decision-making models do not require task specific inputs for model parameterization.

We ask whether two standard models, specifically the diffusion and/or LBA models, can provide a meaningful and coherent account of performance on a complex decision task: a simulated unmanned aerial vehicle (UAV) surveillance task. In particular, we examine how manipulations of the sensory difficulty of the discriminations (i.e., decision uncertainty) and of the level of time pressure (i.e., workload) influence the models' accumulation rate and threshold parameters. We first provide an overview of standard accumulate-to-threshold models, and discuss some recent elaborations aimed at accommodating the effects of multiattribute stimuli and increased decision and response complexity relative to simple choice tasks. We then report the results from an experiment using the UAV task and the accounts of it provided by the diffusion and LBA models. Finally, we discuss the broader implications of our findings and the potential practical applications of accumulate-to-threshold models.

Standard Accumulate-to-Threshold Models

Accumulate-to-threshold models assume that people make decisions by accumulating evidence from the environment until the total evidence for one of the response options reaches a threshold, and the corresponding response is triggered. Although the basic idea of evidence accumulation is simple, complications have been added to model architectures to account for a range of empirical phenomena. Early models of binary choice (e.g., Stone, 1960) assumed the difference in evidence for each choice was accumulated, and that the only source of noise was that evidence varied from moment to moment during accumulation (i.e., a random walk, or its continuous limit, a diffusion process). These assumptions alone were soon shown to be inadequate because they predict that reaction time (RT) distributions for correct and error responses are always the same, which is rarely observed (e.g., Ratcliff & Rouder, 1998). However, the addition of variability among decision trials in the accumulation starting point enabled faster errors to be accommodated, and the addition of variability in the mean amount of evidence available on each trial enabled slower errors to be accommodated (Laming, 1968; Ratcliff, 1978). The current standard version of the diffusion model (e.g., Ratcliff & McKoon, 2008), also adds variability from trial-to-trial in the time to complete nondecision processes.

As shown in Figure 1, the standard diffusion model makes quite specific assumptions about the form of the distributions for each of the three additional types of trial-to-trial variability. The rate of accumulation is assumed to vary from trial-to-trial by a normal distribution with mean v (the "drift rate") and standard deviation, s_v. The start point variability has a uniform distribution with a mean z and width s_z. Finally, Ratcliff and McKoon (2008) assumed that the variability in nondecision processes has a uniform distribution with a mean t_{er} and width s_t. Moment-to-moment, within-trial variability is assumed to have an infinitesimal normal distribution with standard deviation s. In order to make the model identifiable at least one location or scale parameter in one cell of the design to which the model is fit has to be fixed (Donkin, Brown, & Heathcote, 2009); we follow the convention of Ratcliff and colleagues in fixing $s = 0.1$ in all design cells.

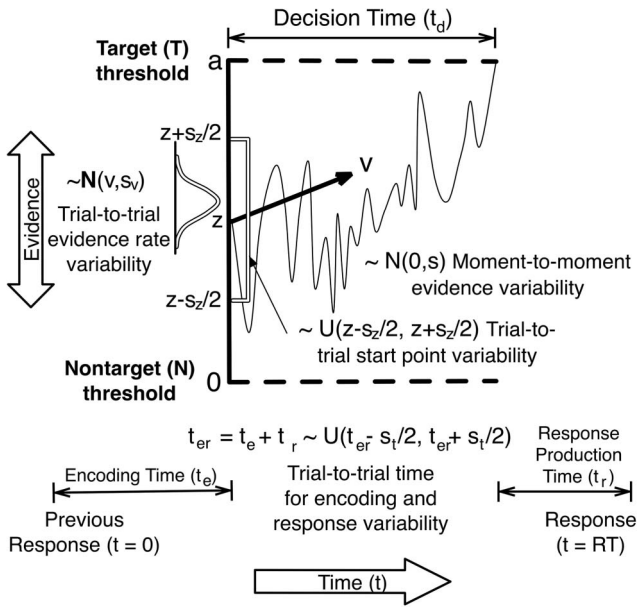

Figure 1. The standard diffusion model applied to the unmanned aerial vehicle task. RT = reaction time.

In previous applications, v has been the primary index of differences across experimental conditions in both evidence strength and quality (i.e., ability to differentiate between choices). Usually s_v is fixed across conditions, although it sometimes varies between stimulus types, for example, words and nonwords in a lexical-decision task (Wagenmakers, Ratcliff, Gomez, & McKoon, 2008), or target and lure words in a recognition memory task (Starns, Ratcliff, & McKoon, 2012). However, s_v also affects evidence quality, which in turn mainly influences accuracy, with the ratio v/s_v being analogous to the measure of discriminability, d', used in SDT.

The response selected by the diffusion model depends on which threshold the evidence total crosses first. In Figure 1 the upper threshold corresponds to a target response and the lower threshold a nontarget response, with the selection of the target threshold illustrated. The magnitude of the upper evidence threshold (a) indexes the overall level of response caution, while the lower threshold is fixed at zero. A larger value of a means more evidence is required for a decision. The mean start point (z) determines response bias by setting the initial level of evidence for each response, with unbiased responding corresponding to $z = a/2$. Differences between the absolute magnitude of rates for stimuli corresponding to different responses can also be seen as instantiating a type of response bias, but a bias in the way evidence is encoded rather than in terms of its initial value.[1] We report the absolute values of drift rates, corresponding to absolute strength of evidence.

The LBA model, illustrated in Figure 2, drops two types of variability in the diffusion model: moment-to-moment noise and nondecision-time noise. The first assumption is a fundamental theoretical difference (see also Brown & Heathcote, 2005), based on the assumption that the effects of trial-to-trial variability are large enough in a relative sense that the effects of moment-to-moment variability can be safely ignored when modeling behavior. The second assumption is based on empirical observations that uniform nondecision-time noise does not usually improve the fit of the model. Under this assumption nondecision time in the LBA is indexed by a single parameter, t_{er}. A second fundamental difference is that there is one accumulator per response, which means that—in contrast to the diffusion model—the LBA is easily extensible to more than two choices. Each accumulator has its own independently sampled normal distribution of evidence accumulation rates, with mean v and standard deviation s_v, and a uniform distribution of start points between a lower bound of zero (again without loss of generality) and an upper bound of A.

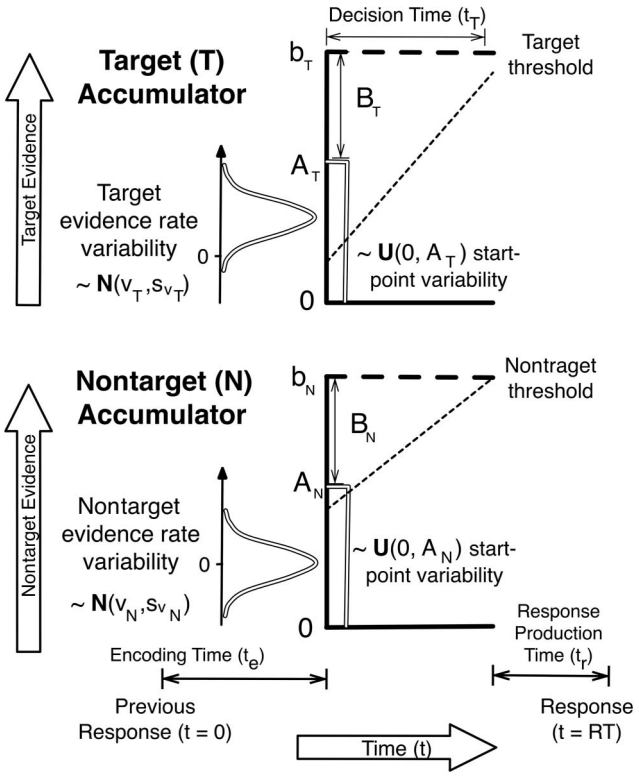

Figure 2. The standard linear ballistic accumulation model applied to the unmanned aerial vehicle task. RT = reaction time.

The response selected corresponds to the LBA accumulator whose evidence total first reaches threshold, which in Figure 2 is the nontarget accumulator. Figure 2 also illustrates how the LBA produces a speed–accuracy trade-off by adjusting thresholds. The initial bias to respond nontarget (i.e., the higher start point for nontarget than target accumulator in Figure 2), which results in the illustrated nontarget response, could have been overcome by a higher threshold. This would allow the faster rate of accumulation for the target accumulator (i.e., a steeper increasing dashed line in Figure 2) to overcome the initial bias. However, as it takes longer for the evidence total to reach a higher threshold, the increase in accuracy due to overcoming the random bias in the start point comes at the price of slower responding. The diffusion model has the same speed–accuracy trade-off mechanism, but also has a second one; the effects of moment-to-moment noise are reduced, and so accuracy is increased by a larger threshold.

Similar mechanisms in the two models also explain adjustments in the relative speeds of error and correct responses. Errors become faster as the level of start-point noise increases relative to the level of trial-to-trial rate noise, and when the distance from the threshold to the maximum level of start point noise is small (e.g., B is close to A in the LBA). Response bias (i.e., a lower threshold for the favored response) causes a pattern of slow errors for the favored response and fast errors for the other response. Stimuli corresponding to the favored response produce slow errors because such errors terminate on the threshold corresponding to the other response, which is further away. In contrast, errors are fast for stimuli corresponding to the nonfavored response, as such errors terminate on the closer threshold.

Because these effects on the relative speed of error and correct responses are quite subtle and specific, they have proved critical for testing the fundamental assumptions of standard evidence accumulation models (e.g., Ratcliff & Rouder, 1998). They have also been used to show accumulate-to-threshold models are superior to alternative accounts of rapid choice (e.g., Wagenmakers et al., 2008). We will examine data from the UAV task to see if similar patterns emerge.

[1] This type of encoding bias is sometime discussed in terms of a "drift criterion," a value above which unidimensional evidence is taken to favor one response (e.g., produce a positive drift rate), and below which it is taken to favor the other response (e.g., produce a negative drift rate).

Complex Tasks

Recently, elaborations of both the diffusion (Diederich, 1997; Fific, Little, & Nosofsky, 2010; Little, Nosofsky, & Denton, 2011; Little, Nosofsky, Donkin, & Denton, 2013) and LBA (Eidels, Donkin, Brown, & Heathcote, 2010; Hawkins et al., 2014; Trueblood, Brown, & Heathcote, 2014) have been proposed to model decisions about multiattribute stimuli, and in the case of the LBA to model more than two response options. These elaborations focus on issues such as whether evidence about attributes is combined (e.g., before accumulation begins, or by attention switches between attributes during accumulation) or not (e.g., separate decisions are made about each attribute then logical rules combine the separate decisions). In both perceptual and higher-level choices (Trueblood, Brown, Heathcote, & Busemeyer, 2013), they have focused on context effects (i.e., the addition of an attribute changing the evaluation of other attributes) arising between multiple attributes when choosing among more than two options. They have also addressed how more than two judgments made about the same set of options, such as choosing both the best and the worst options (Hawkins et al., 2014).

Although these issues are all potentially relevant to applied settings, the elaborated models require more complicated architectures and parameterizations than the standard models. This means they are more demanding in terms of the amount of data they require for reliable estimation than the standard models, which already require more trials than are collected in typical designs. In some cases, the elaborated models are also only estimable in specialized designs, such as the redundant-target and double-factorial paradigms required by Townsend and Nozawa's (1995) systems factorial technology. Further, even though the multiattribute stimuli that we used in our experiment may be considered relatively simple in many applied settings, the mappings of their attributes to target and nontarget categories are, to our knowledge, more complex than anything the elaborated models have been applied to before.

Rather than attempting to develop even more complex or task-specific models and applying them to especially developed designs, we instead asked whether standard models with a flexible parameterization could provide a sufficient approximation to adequately describe data from a detection task within a complex UAV simulation which also included a navigation task (see Figure 3). Our detection task required participants to classify ships moving across a display as either targets or nontargets. This was done by left-clicking the computer mouse within the green (target) or red (nontarget) area of a response box hovering over the ship. Ships could be green, black, or yellow and could have one or more cranes on their deck, or no crane. Nontargets were defined as black or yellow ships with only one crane on deck. Targets were defined as green ships with any deck configuration, or black or yellow ships equipped with zero or two or more cranes. Ships could also have attributes irrelevant to the target classification, including masts and fishing lines.

Although the standard accumulate-to-threshold models do not provide a detailed model of the processing of multiattribute stimuli, they retain a sensible interpretation in the multiattribute setting. In particular, they can be interpreted as being a type of "coactive" model (Little et al., 2013; Miller, 1978, 1982), where evidence about stimulus attributes is combined before accumulation. We hypothesized that under this interpretation the heterogeneity of the target and nontarget stimuli could be accommodated by the model's rate-variability parameters, consistent with its original motivation as accommodating for heterogeneous memorability of words in recognition memory experiments (Ratcliff, 1978).

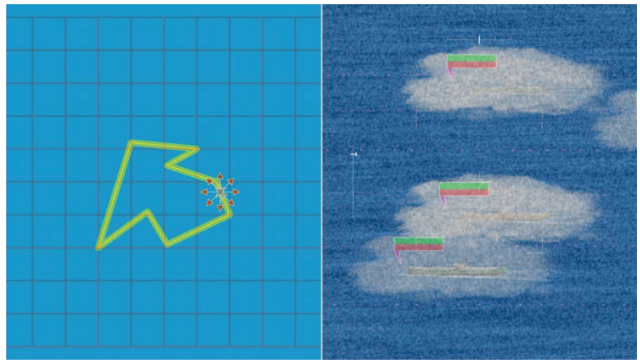

Figure 3. Screenshot of the task display with the navigation task in the left panel and the ocean view used in the detection task in the right panel. See the online article for the color version of this figure.

We sought to understand the psychological implications of standard accumulate-to-threshold models for our complex task by examining selective-influence (see Arnold, Bröder, & Bayen, 2014, for a similar approach applying the diffusion model to simple tasks, and Vuckovic et al., 2014, for an applied example). Selective influence implies that a particular experimental manipulation affects only one type of model parameter and not other types of parameters. In our task we manipulated two factors of key interest in applied settings, time pressure and classification difficulty. We tested selective influence with respect to both factors by fitting models that allowed two different classes of parameters to be affected by difficulty and time pressure. Results from simple choice tasks might lead one to hypothesize that time pressure will affect parameters concerned with the decision process itself, such as the evidence threshold, whereas difficulty will affect parameters concerned with the inputs to the decision process, such as the rate mean and variability. However, as we now discuss, there are reasons to think that selective influence might not be so clean in our more complex task.

Selective-Influence

In our detection task the ocean view had five lanes. Each lane could contain a ship moving across the screen at a constant rate, so that on average a ship was visible for 9 s, and only one ship was visible in a lane at a time. The average number of ships on the screen was constant within a block (i.e., a contiguous set) of trials, but over the course of the experiment the average was manipulated over four levels that increased then decreased twice. As in a real-world UAV scenario, there was greater pressure to respond quickly when more ships were present so that none were missed before they disappeared. We hypothesized that participants would set a lower threshold when more ships were present (i.e., time pressure was higher) in order to speed their decision processes and reduce chances of nonresponse.

With simple tasks, selective threshold influence has mainly been tested through instructions and feedback. Participants are simply

asked to respond either with an emphasis on accuracy or speed, with the instruction usually reinforced by corresponding feedback (i.e., feedback on response accuracy or speed, respectively). However, recent evidence suggests that, for both the diffusion and LBA, the instruction manipulation can also affect accumulation rates (Rae et al., 2014; Starns et al., 2012). Rae, Heathcote, Donkin, Averell, and Brown (2014)demonstrated this to be the case in simple perceptual, lexical, and recognition memory decisions using rigorous model testing procedures, and in the perceptual case with confirmation from nonparametric evidence obtained from a response-deadline paradigm.

The rate changes with speed versus accuracy instructions observed in less complex tasks may occur because participants are not fully engaged when performing a long series of simple decisions, so that there is room for speed instructions to increase rates by increasing attention or arousal. We suspected that our task was much more engaging even at the lowest level of time pressure, so that any increase in attention or arousal as the average number of ships increased would be relatively small. If this were the case, we would observe selective influence on the diffusion and LBA threshold parameters. To test these possibilities, we used the same rigorous model selection tests as Rae et al. (2014), by comparing three types of models. In one, time pressure selectively influences only rate parameters; in a second, it selectively affects only thresholds; and in a third, it can affect both types of parameters (i.e., there is no selective influence). Model selection techniques that take into account both the parsimony of a model and its goodness-of-fit were used to choose the best account of the observed data.

Decision uncertainty occurs in real UAV surveillance tasks when video resolution is compromised by bandwidth limitations between a UAV and its ground control station (McCarley & Wickens, 2005). All of our displays of the ocean view contained dynamic pixilation, and uncertainty was also manipulated over three levels using clouds that constantly obscured the ship stimuli to varying degrees (see Figure 3). In simple tasks, perceptual manipulations of the difficulty of the target versus nontarget discrimination, such as our manipulation of cloud opacity, are assumed to selectively influence parameters related to the rate of evidence accumulation. In the diffusion model perceptual manipulations are most often assumed to affect the mean accumulation rate parameter, with the rate decreasing as difficulty increases. In the LBA they are assumed to affect the difference between the rate for the accumulator that matches the stimulus (e.g., the target accumulator for a target stimulus) and the rate for the accumulator that mismatches the stimulus (e.g., the nontarget accumulator for a target stimulus); as difficulty increases the difference decreases. The diffusion model can be thought of as using this difference as the input to accumulation in the manner of a coactive model, hence the similar effect of its rate parameter and the rate difference in the LBA. In general, accumulator models such as the LBA are more flexible in this regard, in that an increased rate difference need not necessarily increase response speed, but an increase in the diffusion's rate parameter must do so, all other things being equal.

In contrast to rate parameters, accumulator parameters, such as threshold and level of start-point noise, are usually assumed to be unaffected by classification difficulty. This is because thresholds are thought to change relatively slowly, usually due to strategic factors, and so are not immediately influenced by the stimulus on the current trial. Similarly, start-point noise is assumed to be determined by accumulation occurring before the stimulus appears (e.g., Laming's, 1968, "premature sampling") or persistent bias effects caused by previous decisions (e.g., when evidence has not returned to baseline between trials, Brown, Marley, Donkin, & Heathcote, 2008). In particular, it is considered circular to adjust properties of the accumulation process based on knowledge of stimulus properties that are the focus of the classification decision (e.g., whether it is a target or nontarget), as this knowledge is only obtained once the accumulation process terminates. Thus, if difficulty is manipulated through a property integral to the classification—such as the degree to which stimulus brightness departs from a middle value in a bright versus dull classification task (Ratcliff & Rouder, 1998)—it would not make sense to allow thresholds to vary with difficulty.

In our task, in contrast, the difficulty manipulation, cloud opacity, is not integral to the target versus nontarget classification. If participants can discriminate different levels of cloud opacity, it might be to their strategic advantage to set lower thresholds for easier stimuli. This is because there is a nonlinear relationship between accuracy and the threshold, so that when accuracy is high, reducing the threshold can markedly improve speed with only a small decrement in accuracy, whereas when accuracy is low increasing the threshold can markedly improve accuracy while only slowing responding slightly (e.g., Ratcliff & Rouder, 1998). Given that time pressure places a premium on fast processing it would be desirable to set a lower threshold for easier stimuli (i.e., those with less opaque clouds). Such flexible threshold adjustment has been observed in task-switching paradigms, where the threshold is reduced when a prestimulus cue indicates that the upcoming trial would be easier (Karayanidis et al., 2009), although older participants are less flexible (Whitson et al., 2014). However, any improvement in speed for easy stimuli in our task might be offset if threshold adjustment takes some time to achieve and so delays the onset of accumulation, or if attention capacity devoted to threshold adjustment decreases the rate of evidence accumulation (see Schmiedek et al., 2007, for a link between capacity and rates).

To test these different possibilities, we again compared three types of models, in this case allowing different types of parameters to be affected by difficulty. In one, we made the conventional assumption, that difficulty selectively influences only rate parameters. In the second, difficulty selectively affects only thresholds, and in the third there is no selective influence, so both types of parameters can be affected by difficulty. We examined all possible combinations of these models and the corresponding models concerning the influence of time pressure, and again used model selection based on both parsimony and goodness-of-fit to choose the best combined account.

Serial Versus Parallel Processing

The presence of multiple stimuli on the screen raises the question of whether participants make decisions in series or in parallel. This question of "cognitive architecture" is intimately related to the question of capacity. Capacity in evidence accumulation models is operationalized through the rate parameter (e.g., Eidels et al., 2010), with reduced capacity causing a decreased rate of accumulation. In a parallel architecture, participants might initiate a separate decision process (i.e., a single accumulator for the diffusion or a pair for the LBA) for each lane as soon as a ship appears in

it. If capacity is limited then the rates for each decision process will slow as more processes are added, which in turn will slow decisions, and so increase the chance of missed responses. In a serial architecture, only one decision process runs at a time, with a new process being initiated after the last process reaches threshold. In the serial case all available capacity can be dedicated to the active decision process, but misses may still occur because some stimuli disappear from the screen before a decision process can be activated for them. Although capacity is often viewed as fixed, it may be changed by effort, which can be viewed as a form of mental resource related to capacity (Kahneman, 1973). Thus, changes in available capacity may occur due to changes in cognitive effort, which affects the allocation of the resources to the experiment-defined task, or by an increase in arousal, which affects the availability of resources (Humphreys & Revelle, 1984).

Serial and parallel processing architectures are remarkably hard to differentiate when capacity can also vary (Townsend, 1990). For example, it might be thought that a slowing of processing as the number of stimuli increases is indicative of serial processing, but this need not occur in a serial architecture if capacity also increases with the number of stimuli (so called "super-capacity"). On the flip side, a parallel architecture can result in slowing if there is a limited amount of capacity that must be shared, so that each process slows as the number of items increases. Systems factorial technology (Townsend & Nozawa, 1995) provides rigorous nonparametric tests of these issues based on the full distribution of RT in specialized and highly controlled designs.

In our less controlled design we sought to address the question of architecture in a f, by scoring RT for a decision about a ship either from the time it appeared on the screen (absolute scoring) or from the time that the preceding decision was made (relative scoring). Absolute scoring corresponds to a parallel processing assumption, where a decision process starts accumulation some time (equal to the encoding time) after the ship appears, and terminates some time (equal to response-production time) before a response to that ship. Relative scoring corresponds to a serial processing assumption, where accumulation for each ship begins only after accumulation for another ship terminates.

The correspondence is approximate for both types of scoring. In the parallel case, if a process terminates while response production for a previously terminating process is occurring, its response will likely be delayed. However, it seems likely that such occurrences may not happen often, and that in any case their primary effect will be to increase the nondecision time estimate without much affecting other parameters, or model fit (see Hawkins et al., 2014). In the serial case, accumulation may begin during the response production time for the previous process, but again this seems likely to only have an effect on the nondecision time estimate, in this case causing an underestimate. A greater challenge in the serial case is *partial processing*, which could occur if a ship that is the current focus disappears and participants do not make a response to it. Relative scoring will then cause the apparent RT for the next decision to be increased (as it will include the time spent on the ship that disappeared) and to be more variable (as the delay could vary from negligible to slightly less than the several seconds that it takes to make a decision), so that more slow responses will be recorded than predicted by the accumulate-to-threshold model. That is, frequent partial processing should cause the serial model to underestimate the right RT distribution tail.

Model fit should, however, be much more affected if the wrong type of scoring is applied. Suppose processing is parallel, but relative scoring is applied. In that case scored RT will be less than true RT on every trial by a randomly varying amount. If processing is serial and absolute scoring applied then scored RT will be greater than the true RT by a randomly varying amount on every trial. All other things being equal the additional random variation should make the scored RT distribution more symmetrical than the true RT distribution. Both the diffusion (Ratcliff, 2002) and LBA (Heathcote, Wagenmakers, & Brown, 2014) are unable to fit RT distributions that are symmetric, so a parallel architecture will be supported if model fit is better for absolute scoring, and a serial architecture will be supported if model fit is better for relative scoring.

Experiment

In the forgoing we outlined a strategy for testing the suitability of the LBA and diffusion models for modeling the UAV surveillance task. We now specify this strategy in more detail. The first step involves fitting sets of LBA and diffusion models to both the absolutely and relatively scored data using the maximum likelihood methods described in Donkin, Brown, and Heathcote (2011). We will select a parallel architecture if the misfit to the absolutely scored data is least, and the serial architecture if the misfit to the relatively scored data is least. We quantified misfit by the deviance statistic (D), which is proportional to the negative of the maximized likelihood, and analogous to the familiar sum of squared errors measure of model misfit used in the general linear model. Specifically, the deviance equals -2 times the maximized log-likelihood. The -2 multiplier means that models with the smallest deviance fit better, and that differences in deviance between nested models (i.e., models with different numbers of parameters where the model with fewer parameters is a simplified version of the model with more parameters) are approximately distributed as $\chi^2(p_1\text{-}p_0)$, where p_1 and p_0 are the number of parameters in the nesting and nested models, respectively. Note that the magnitude of the deviance is not meaningful, as it varies with the units used to define time, and has no natural origin.

Subsequent analysis will focus on the selected (serial or parallel) LBA and diffusion model sets, and the data obtained by the corresponding scoring method (i.e., absolute scoring for parallel and relative scoring for serial). Variants of the LBA and diffusion models within each set instantiate different selective-influence assumptions, and have different numbers of parameters. Hence variants have different degrees of flexibility to fit the data, and also to potentially over fit it (i.e., accommodate chance patterns). We used the Akaike Information Criterion (AIC, see Burnham & Anderson, 2004)[2] to select the variant that provided the best tradeoff between flexibility and goodness-of-fit. AIC $=$ D $+ 2p$,

[2] We also examined the commonly used alternative Bayesian Information Criterion, BIC $=$ D $+ p \ln(N)$, where N is the number of data points. Because the BIC complexity penalty is much larger (as $\ln(N) \gg 2$), BIC prefers simpler models than AIC. In a Bayesian framework, assuming a unit information prior, BIC has asymptotically consistent model selection, but in nonasymptotic samples BIC tends to only allow for the large effects, ignoring smaller but still reliable differences. We found this to be clearly the case in our data, with BIC-based selection only allowing for stimulus effects. As it was clear that workload and difficulty effects are reliable, we preferred to use AIC.

where p is the number of model parameters. The $2p$ term acts as a complexity penalty, disadvantaging models with more parameters unless each extra parameter decreases deviance by more than 2. Note that, in contrast to the deviance, AIC can be used to compare non-nested models, either within a set of variants of the same model or across qualitatively different models such as the LBA and diffusion models.

We tested our selective influence hypothesis by examining which selective influences, if any, were retained by AIC in both the diffusion and LBA models. Note that our main aim was not to determine whether the diffusion model was better than the LBA or vise versa, although this determination can be made using AIC. Rather, we sought to investigate whether both models supported the same conclusions, which would indicate that these conclusions did not strongly depend on the different detailed assumptions of the two models.

Our final set of analyses focused on the AIC of the selected models. We examined their goodness-of-fit graphically, to determine whether they captured all of the effects of the experimental manipulations on both accuracy and RT distribution. We quantified RT distribution for correct and error responses, using the 10th percentile (reflecting fast responses), the 50th percentile (the median) and the 90th percentile (reflecting slow responses). Although some misfit is to be expected and even desirable (otherwise one might suspect over fitting) parameters cannot provide a valid distillation of psychological causes if they do not come from a model that captures the phenomena of interest, such as workload and difficulty effects. We then examined the estimated parameter values in order to gain a quantitative psychological understanding of the effects of workload, difficulty and differences between target and nontarget processing.

Model Variants

The sets of models that we fit were generated from a highly flexible "top" model. The other models in the set are simpler variants of the top model that impose successively more stringent selective-influence assumptions. The simplest model in the set only has the flexibility to perform better than chance and to display response bias, but otherwise has the same parameters across all experimental conditions. For the diffusion model, the simplest variant estimates a single value for each of its seven parameters (a, z, s_v, v, s_v, t_{er}, s_t). For the LBA, the simplest variant estimates a single value for three parameters (A, s_v, t_{er}) and two values each for the v and B parameters, one corresponding to each accumulator.

The parameters of the LBA are allowed to vary as a function of two accumulator-related factors, a *response* (R) factor and a *match* (M) factor in order to account for response bias and accuracy respectively. The LBA accounts for response bias by allowing the threshold (B) to vary across the accumulators for the different responses. For example, a lower value of B for the target accumulator causes a target bias. This is captured by the response factor, which has two levels (target accumulator vs. nontarget accumulator). The LBA accounts for accuracy by allowing the rate (v) to be greater when the accumulator matches the stimulus than when it mismatches. If performance is greater than chance, then for the target accumulator, v will be greater for targets than nontargets, while for the nontarget accumulator, v will be greater for nontargets than targets. This is captured by the match factor, which has two levels (match vs. mismatch). In more complex models it can also make sense to allow the A parameter to be a function of the response factor (as like B it is a property of an accumulator) and the s_v parameter to be a function of the match factor (as like v it is a property of the stimulus).

In Donkin et al.'s (2011) method there is an exponential increase in the number of models in the set as the number of factors (NF) that are allowed to affect the parameters of the top model increases (2^{NF}). Our new domain of application required us to use a large value of NF in order to test rather than assume a range of conventional selective influence assumptions. Although the computational resources dedicated to this project were large (many weeks of fitting on several 64 core servers), the exponential increase meant that we could not consider a top model where all experimental factors affect all parameters. There were three experimental factors: stimulus (S, target vs. nontarget), workload (WL, four rates at which ships crossed the screen), and difficulty (D, three levels of cloud opacity). If, for example, all three were allowed to affect every one of the diffusion model's seven parameters the set would have $2^{7 \times 3}$ variants, requiring fitting of over two million models for each participant.

Fortunately there is a strong case for assuming that accumulator-related parameters cannot be influenced by stimulus-related factors that are unknown at the start of the trial, which in our case is the stimulus (S) factor. This is because, as previously discussed, it is the outcome of the accumulation process that determines the stimulus classification, so it would be circular to assume the accumulator settings (e.g., a, z, and s_z in the diffusion model) determining the outcome can be influenced by that outcome. Even so, the reduction in the set size was not sufficient (e.g., for the diffusion $2^{4 \times 3 + 3 \times 2} = 262{,}144$ models per participant).

Hence, we assumed further restrictions, estimating only a single value applying to all conditions for the nondecision time (t_{er} for both LBA and diffusion, and s_t as well for the diffusion) and bias variability (A for the LBA and s_z for the diffusion) parameters. These assumptions are not necessary (e.g., it is possible that encoding time, and hence t_{er}, may differ depending on difficulty, and that response bias may change with workload), but do result in a reasonable number of models per participant, $2^9 = 512$ for the diffusion, and $2^{10} = 1{,}024$ for the LBA. The extra factor in the LBA occurred because we allowed s_v to vary with the match (M) factor in the top model. In previous LBA applications greater mismatch than match variability has been consistently found (e.g., Heathcote & Hayes, 2012; Heathcote & Love, 2012; Rae et al., 2014). Note that, should these restrictions be wrong, we might expect clear misfit of the data.

In summary, the top diffusion model has 67 parameters and the top LBA Model 121 parameters. The much larger number for the top LBA model is because its extra match-factor effectively doubles the number of v parameters in its top model (48, $S = 2 \times D = 3 \times WL = 4 \times M = 2$) relative to the diffusion. For s_v in the LBA the number is one less (47) as one parameter must have a fixed value in order make the model identifiable (Donkin, Brown, & Heathcote, 2009). The remaining 26 LBA parameters include 24 for B ($R = 2 \times WL = 4 \times D = 3$) and one each for A and t_{er}. For the diffusion there are 24 parameters each for v and s_v ($S = 2 \times D = 3 \times WL = 4$), 12 each for a ($WL = 4 \times D = 3$), four for z ($WL = 4$), and one each for t_{er}, s_t and s_z.

The large numbers of parameters in the top models mean that optimization algorithms usually fail to find best-fitting parameter values unless search starts from reasonably good guess. In Donkin et al.'s (2011) method this difficulty is overcome by first fitting the simplest model, then using the best fitting values to start search for all models with one extra factor, and then repeating this process up to the top model. In this way more complex models are fit from multiple start points given by all models that have one less factor (e.g., eight different start points for both the top LBA and diffusion models). Although this increases the number of fits required (e.g., 2,305 for the 512 diffusion variants) it also greatly reduces the likelihood of finding suboptimal solutions.

Method

Participants

Thirty-two participants took part in the study, consisting of 22 females and 10 males, with a mean age of 21.59 years (SD = 4.38). They were undergraduate students from the University of Queensland, Brisbane, who received course credit for their participation, and paid participants who were compensated $10 for their time.

Experimental Task

The UAV simulation comprised of a UAV navigation task and a target detection task using a split-screen configuration (see Figure 3), reflecting the competing tasks personnel often perform in complex work systems. Interactions and inputs with the target detection task and navigation task were independent; participants could only interact with one task using the computer mouse at any given point in time.

In the navigation task participants had to guide a UAV (traveling at 20.6m/s) around a predefined polygon-shaped path, maintaining the UAV within a 200-m boundary from the center of the path (see left panel, Figure 3). If the UAV exited the path boundary an alarm sounded until the UAV was navigated back within the boundary. The participant navigated the UAV using eight heading-control navigational arrows. The heading change in the UAV's direction was gradual, rather than instantaneous. The path geometry was designed so that occasional interventions were required from the participant to keep the UAV on-track.[3] Note that a navigation correction was counted as a decision for the purposes of relative scoring.

The detection task simulated a camera ocean view, with multiple ships passing the camera's field of view. The simulation included five ship lanes, allowing up to five ships to appear simultaneously, as no more than one ship could appear in a lane at any one time. The number of ships present at any given point in time varied. Participants were required to classify ships as targets or nontargets depending on the ship's configuration, which varied with respect to hull color (green, black, and yellow) and equipment on deck. Each of three deck positions could contain a crane, a mast, or no equipment. Of the 27 possible equipment arrangements, a subset of 16 was utilized. This subset comprised all 12 arrangements that included one and only one crane, all three arrangements that included two cranes and one mast, and the one arrangement that included three masts.

Regardless of deck equipment, green ships were always targets. Black and yellow ships were targets if they had zero or two cranes with any number of masts, and were nontargets if they had one crane with any number of masts. On each trial each deck configuration had an equal probability of being selected, as did each hull color, so targets appeared on 50% of trials on average. These settings simulated a situation in which target and nontarget categories were quite heterogeneous, as can often be the case in applied contexts.

Participants classified a ship by left-clicking the computer mouse on a response box that hovered above the ship as it traveled across the screen; clicking the red and green portion of the response box corresponded to making a nontarget classification and a target classification, respectively. Participants were able to change their responses by left-clicking the alternative choice in the response box, and their last response was recorded as their ship classification. To simulate the characteristics of surveillance footage, noise was introduced to the simulated ocean view by adding a dynamic gray-scale noise overlay.

Experimental Design

The design had 24 conditions created by fully crossing three within-subjects factors: stimulus (target and nontarget), time pressure (low, medium, high, and very high), and discrimination difficulty (easy, medium, and hard). Participants completed 16 blocks, each lasting 3 min. The blocks followed a ramp-up, ramp-down sequence in terms of time pressure.[4] Participants experienced an increase in time pressure, and then a decrease in time pressure, and this sequence was then repeated. The difficulty manipulation was randomized within blocks. The time pressure and difficulty levels were calibrated, based on pilot studies with the same population, to produce significant differences in accuracy.

The time pressure manipulation varied intership onset times. The ships entered the screen more frequently as time pressure increased, thereby reducing the time available to attend to individual ships. The manipulation was implemented by varying the initial distance of ships from the simulation field of view while keeping ship speed constant at 25.72m/s. The initial distance of each ship was chosen randomly from a uniform distribution ranging from 180 m to 730 m, 580 m, 430 m, or 280 m for the low, medium, high, and very high time pressure conditions, respectively. For the participants included in the analyses, the average numbers of ships in the low, medium, high, and very high time pressure trials were 37.5, 42.1, 48.9, and 57.9 per trial, respectively (SDs = 1.74, 1.35, 1.64, 1.10, respectively).

The discrimination difficulty manipulation varied the transparency of cloud texture superimposed on the ships. The superim-

[3] The median time between alarms was 15 s, with the exact time depending on how effective the previous renavigation was (i.e., intervals were longer after better corrections). Participants mostly responded quickly to the alarm (median time 2 s) but occasionally allowed the alarm to sound for longer periods.

[4] In subsidiary analyses we did not find any indications of interactions between the effects on speed and accuracy reported in the main body of the paper and factors related to the blocked changes, phase (ascending vs. descending) and cycle (1st vs. 2nd). However, division by these factors may have led to a lack of power to detect more subtle effects.

posed cloud moved with the ship, constantly obscuring the ship features. Lower difficulty corresponded to lesser cloud opacity, thus making it easier to discern ship attributes. Cloud opacity was 65%, 82.5%, and 100% of that of a base cloud image for easy, medium, and hard difficulty, respectively. Each transparency condition occurred with equal probability.

Procedure

Participants viewed an audiovisual presentation that explained how to perform the task, and then completed a training phase of four blocks lasting 90 s each. Time pressure increased from low to very high over the four blocks. Throughout training, participants could review a printout that explained the ship classification scheme. After training, participants were asked by the experimenter if they understood the task and had the summary printout removed from their desk. Once any outstanding issues were clarified by the experimenter and the participant reported that he or she understood the tasks, the experimental phase of the study began. In total, practice and experimental blocks took 1 hr to complete.

Results

Although the majority of participants responded to most stimuli, a few had many missing responses. Because this resulted insufficient responses for stable model fitting, we eliminated the data from three participants with more than 50% missing data (58%, 59%, and 72%). When we repeated our analyses with a stricter criterion of less than 25% missing values—eliminating four more participants (33%, 42%, 42%, and 49%)—we reached essentially the same conclusions, so we present results for the larger sample of 29 participants.

We analyzed nonresponse rates with generalized linear mixed models with a probit link function, using the R package *lme4* (Bates, Maechler, Bolker, & Walker, 2014), with inferences conducted via Wald's chi-square tests, as implemented by the R *car* package (Fox & Weisberg, 2011). All experimental factors had strong main effects, with more nonresponses for nontargets (11%) than targets (9%), $\chi^2(1) = 20.78$, $p < .001$, and increasing nonresponse rates as difficulty increased (7%, 10%, 14%), $\chi^2(2) = 219.21$, $p < .001$, and as workload increased (8%, 9%, 11%, 12%), $\chi^2(3) = 65.63$, $p < .001$. However, interactions between these effects did not approach significance ($ps > .321$).

Figure 4 plots the distributions of raw RTs scored by absolute and sequential scoring methods. RTs obtained by relative scoring show the strong positive skew characteristic of simple choice tasks (e.g., Luce, 1986). RTs obtained by absolute scoring, in contrast, are much more symmetric. The symmetry of the absolute scoring distribution provides preliminary evidence against the parallel model, whereas the positive skew of the relative scoring distribution supports the serial architecture.

For further analyses, including the model fits, 0.75% of sequentially scored responses, and 1.26% of absolutely scored responses, were censored due to RTs outside the range 0.25 s to 8 s, with at most 4.4% removed from any one individual. The lower cutoff is fairly conventional for simple tasks, because below it there is insufficient time for stimulus-based choice responses to be generated (Luce, 1986), so at least that time would be required for the more complex decision required here. The upper cutoff is several times larger than typically assumed for simple tasks. We

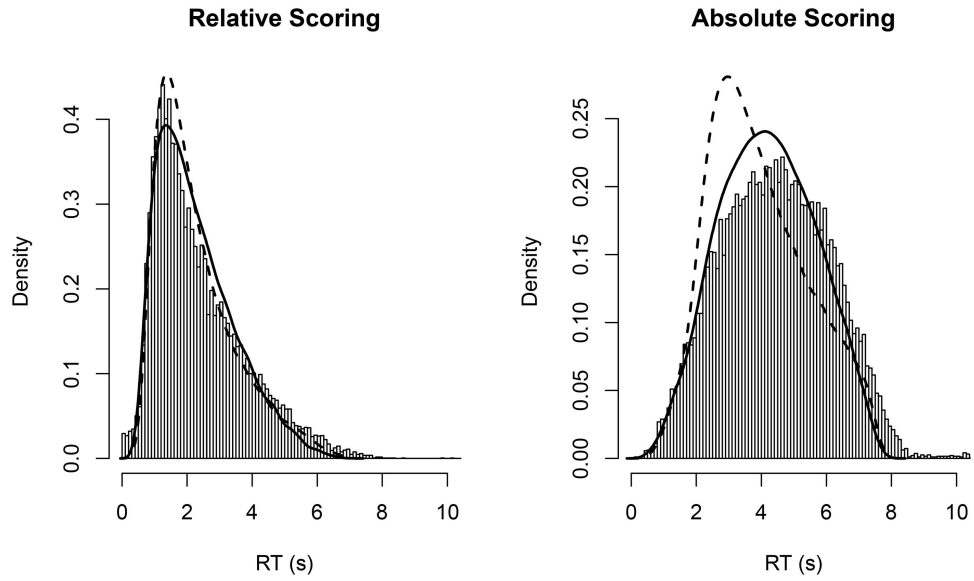

Figure 4. Raw reaction times (RTs) obtained by either relative scoring (time since the last response) or absolutely (time since the ship appeared on the screen). RTs from all experimental conditions for both correct and error responses are plotted in a single histogram with 0.1 s bins. Solid lines are the density corresponding to the RTs predicted by the top linear ballistic accumulation model and dashed lines the density for the top diffusion model. Predicted RTs were obtained from order statistics of the model corresponding to data order statistics for each design cell and for correct and error responses separately. The R density function with default settings was then used to calculate predicted densities for the aggregated predicted data.

used the same cutoff for absolutely scored RTs so that we could compare the two models on the same set of RTs. For the sequentially scored case the upper cutoff may be overly long as it removes hardly any responses (0.04%). However, we were interested in examining the degree to which partial processing produces long RT; if it plays more than a minor role then the serial model should underestimate the long right tail of the RT distribution evident in Figure 4.

Table 1 and 2 show, for the diffusion and LBA models respectively, the deviance and AIC results, summed over participants, for both the top variant and the AIC selected variant (i.e., the variant whose AIC value summed over participants was the lowest of any variant). The tables show very strong support for the serial architecture (i.e., much better fits to the data scored relatively compared with absolutely) for both diffusion and LBA models. Figure 4 plots the fit of the top models to a histogram both types of data; the inability of both the diffusion and LBA model to fit the absolute RT data is evident, whereas both capture the shape of the distribution of the relative RT data quite well. Hence, we limit further consideration to the relative RTs corresponding to the serial architecture. Before considering the modeling results in more detail, we report the results of a conventional analysis of the relative RT data.

Error Rates and RT

We used generalized linear mixed models with a probit link function to analyze error rates (ER), and general linear mixed models to analyze four RT-based measures, the mean and standard deviation of participant's RT for correct and error responses. All analyses were carried again using *lme4* and the *car package* for Wald's chi-square tests. All five analyses revealed a consistent pattern of significant effects for the main effect of workload and an interaction between stimulus and difficulty (see Table 3). In most cases the stimulus and difficulty main effects were also significant, but with no other interactions achieving significance.

Figure 5 illustrates these results for the measures that are conventionally examined, error rates and mean RT for correct responses, as the same general trends were evident in error RT and both correct and error RT standard deviations, except that the effect of difficulty on nontargets was not significant ($ps > .3$), with small trends toward decreases with difficulty for the error measures.

Although both errors and correct RT had a difficulty by stimulus interaction, their causes were very different. Errors increased as difficulty increased, however, the effect was stronger for targets than nontargets. On the other hand, there was a strong decrease in correct RT for nontargets as difficulty increased, which was accompanied by a slight increase in correct RT for targets. Thus, there was a double dissociation for the error rate and correct RT, with greater difficulty mainly increasing target error rates, whereas for RT, greater difficulty mainly decreased nontarget RT.

Workload also exerted the opposite effect on error rates and mean correct response times. Increases in workload produced higher error rates but faster mean correct RTs, especially for the two highest workload levels. These opposing effects are clearly indicative of a tradeoff, whereby participants scarified accuracy for speed as workload increased. Although accuracy is lost, this strategy has a clear benefit; faster responses under higher workloads reduce the likelihood of missed responses.

Model Selection and Fits

Tables 1 and 2 show clear and consistent support in both the diffusion and LBA models for a selective influence of workload on threshold parameters (B in the LBA and a in the diffusion), as the WL factor is dropped from all rate parameters in the AIC-selected variant. Similarly, there is clear support that is consistent between models for a selective influence of difficulty on the mean rate (v), with the effect of difficulty dropped from the threshold parameters for both models. The two models are also consistent in that they both drop the effect of difficulty on rate variability (s_v) in the AIC-selected variants, although they both retain an effect of stimulus on rates means and standard deviations. AIC-selected variants have markedly fewer parameters than the top models but there is little decrement in fit; the reduction in fit for the AIC-selected relative to top models did not approach significance for the diffusion, $\chi^2(1479) = 991$, $p > .999$, or the LBA, $\chi^2(2784) = 264$, $p > .999$. Hence, in further analyses we focus on the AIC-selected models.

Figures 6 and 7 represent, respectively, the goodness-of-fit of the AIC-selected diffusion and LBA models. Tables 1 and 2 indicate that the LBA provides a better account of the data, both in absolute sense (deviance) and when adjusted for the number of estimated parameters (AIC). However, Figures 6 and 7 show the two models provide a fairly similar descriptive account in terms of error rates and RT distribution. Figure 8 focuses on how well the models are able to account for the interaction between the speed of correct and error responses and stimulus (error bars, which are large for error RT, are omitted for clarity).

Clearly the model-based account of accuracy is excellent, with very little misfit for either model. The fit for correct RT distribution also captures most of the qualitative trends in the data, but underestimates some effects quantitatively. The clearest issue is with underestimation of the slowest (90th percentile) responses, particularly for nontargets at lower workloads. There is also a tendency for both models to underestimate the size of the workload

Table 1
Diffusion Fits Summed Over Participants

Scoring	Variant	a	v	s_v	z	p	Deviance	AIC
Relative	Top	WL,D	S,D,WL	S,D,WL	WL	67	76705	80591
	AIC	WL	S,D	S	—	16	77696	78624
Absolute	Top	WL,D	S,D,WL	S,D,WL	WL	67	93107	96903
	AIC	D	S,D	—	—	14	93896	94708

Note. AIC = Akaike Information Criterion; p = number of parameters per participant. Model factors are WL = workload; D = difficulty; S= Stimulus.

Table 2
Linear Ballistic Accumulation Fits Summed Over Participants

Scoring	Variant	*B*	*v*	s_v	*p*	Deviance	AIC
Relative	Top	R,WL,D	S,D,WL,M	S,D,WL,M	121	74986	82004
	AIC	R,WL	S,D,M	S	25	75250	76700
Absolute	Top	R,WL,D	S,D,WL,M	S,D,WL,M	121	83194	90212
	AIC	R,D	S,D,M	S	21	83652	84986

Note. AIC = Akaike Information Criterion; *p* = number of parameters per participant. Model factors are WL = workload; D = difficulty; S= Stimulus. Accumulator factors are M = match; R = response.

effect, and to a lesser degree the difficulty effect (including the interaction with correct vs. error response) on RT.[5] Below we examine in more detail the model's quantitative account of these effects, both on average and in individual participants. However, we first examine model parameter estimates in order to understand what drives their ability to provide a good qualitative account of these effects.

Effects on Parameters

We analyzed parameters from the AIC-selected diffusion and LBA models using general linear mixed models, first addressing the accumulator parameters for both, and then the rate-related parameters. There was a significant decrease in the diffusion threshold (*a*) parameter as workload increased, which is illustrated in Figure 9. There was also a trend for a target bias, as indicated by a start-point to threshold ratio (*z*/*a*) of 0.49 (we assumed the lower boundary corresponded to a target), but this was not significantly different from unbiased (i.e., 0.5, $t(28) = 1.51$, $p = .143$).

For the LBA there was a main effect of the match, $\chi^2(1) = 6.78$, $p = .009$, and workload, $\chi^2(3) = 15.96$, $p < .001$, factors, but no interaction, $\chi^2(3) = 0.64$, $p = .888$. The match main effect was due to a higher threshold for the nontarget than target accumulator (2.14 vs. 1.79, i.e., a target bias). Figure 9 illustrates the significant decrease in the LBA threshold (*B*) parameter with workload. Thus, although differing quantitatively, the two models agree that participants were slightly target biased, and decreased their evidence threshold markedly as workload increased.

For the diffusion drift rate, there were significant main effects of stimulus, $\chi^2(1) = 19.83$, $p < .001$, difficulty, $\chi^2(2) = 6.31$, $p = .043$, and their interaction, $\chi^2(2) = 6.58$, $p = .037$. As illustrated in Figure 10, the interaction occurred because rates decreased strongly with difficulty for targets, but not for nontargets. For LBA mean rates, there was a significant main effect of stimulus, $\chi^2(1) = 5.45$, $p = .020$. The drift rate was greater for targets than nontargets (2.21 vs. 1.92). The only other significant main effect was of the match factor, $\chi^2(1) = 137.94$, $p < .001$. The mean rate was greater for the matching than the mismatching accumulator (2.79 vs. 1.34), which is to be expected, given that performance was above chance. The only remaining effects to achieve significance were two interactions involving the match factor with the stimulus factor, $\chi^2(1) = 29.27$, $p < .001$, and the difficulty factor, $\chi^2(2) = 6.91$, $p = .032$. As illustrated in Figure 10, this was because the difference between the rates of the matching and mismatching accumulators was much larger for targets than nontargets (2.11 vs. 0.78) and decreased with increased difficulty (1.81, 1.49, and 1.03).

Figure 11 shows a trend for the diffusion rate standard deviation (s_v) to be greater for targets than nontargets, but this did not achieve significance, $\chi^2(1) = 1.11$, $p = .293$. For the LBA rate standard deviation, in contrast, there was a significant interaction between the match and stimulus factors, $\chi^2(1) = 7.04$, $p = .008$, with the match standard deviation for targets greater than for nontargets, whereas a trend for the opposite difference in nontargets. We explored whether these patterns of results could be explained by the heterogeneous difficulty of target ship definitions. As reported in the appendix of this paper, when we investigated more closely we found that there were marked differences in performance for "easy" targets, which were defined by a single feature (a green hull), and "hard" targets, which were defined by a combination of features. The reduced sample size associated with this further breakdown of the data made these analyses less stable, but they were generally consistent with the results presented here.

Individual Differences in Effects

Our previous analyses indicate that, for both error rates and mean RTs, there was a strong main effect of workload and an interaction between difficulty and stimulus effects. To quantify individual differences we calculated a number of difference scores for each individual participant's data, and for each individual participant's model fit. The first score was the difference between the highest and lowest workload conditions. Second, we calculated the difference between the lowest and highest difficulty conditions, but separately for targets and nontargets in order to take account of the interaction. Error rates also had a stimulus main effect, so for them we also calculated a nontarget minus target difference score. For mean RT the stimulus effect interacted with the accuracy of the responses, so we calculated the nontarget minus target difference separately for correct and error responses. In each case these calculations were done separately for each participant. Tables 4 and 5 report these differences, averaged over participants. They also report the correlations of the individual participant's differences in the data with the corresponding differences for the AIC-selected diffusion and LBA models.

Table 4 shows that for error rate, in agreement with the good fits in Figures 6 and 7, on average both models provide a quite

[5] A reviewer pointed out that although the LBA did a better job at capturing the overall level of correct and error RT it was less able to capture the interaction with stimulus, particularly for targets. It is possible this was related to the heterogeneity of difficult among targets examined in the appendix, but sample sizes were so reduced in that analysis that it was difficult to discern consistent trends in the relatively rare error responses.

Table 3
Analysis of Variance Results for Conventional Moment-Based Performance Measures

		Error rate		Mean correct RT		Mean error RT		*SD* correct RT		*SD* error RT	
Effect	*df*	χ^2	*p*	χ^2	*p*	χ^2	*p*	χ^2	*p*	χ^2	*p*
WL	3	15.83	<.001	96.34	<.001	84.31	<.001	12.56	<.001	16.29	<.001
S	1	434.91	<.001	67.53	<.001	8.85	.003	24.56	<.006	1.17	.280
D	2	185.60	<.001	20.79	<.001	30.36	<.001	2.87	.238	12.75	.002
S × D	2	130.18	<.001	57.60	<.001	8.29	.016	6.21	.045	6.27	.044

Note. RT = reaction time; WL = workload; D = difficulty; S = Stimulus.

accurate account the sizes of the differences related to stimulus and difficulty. In each case there is also a very strong and highly significant correlation between the data and model differences. For the workload effect there is some underestimation, but that effect is quite small in the data (see also Figure 5) and the small range of this difference explains the weak correlations.

Table 5 shows that, again in agreement with Figures 6 and 7, the account of RT effects is not as good as that of errors, as there is a pervasive tendency for the models to underestimate the sizes of the differences. However, the correlations between data and model differences remain substantial, and in most cases highly significant.

Discussion

In this paper we applied accumulate-to-threshold models as dynamic measurement models that, like signal-detection theory,

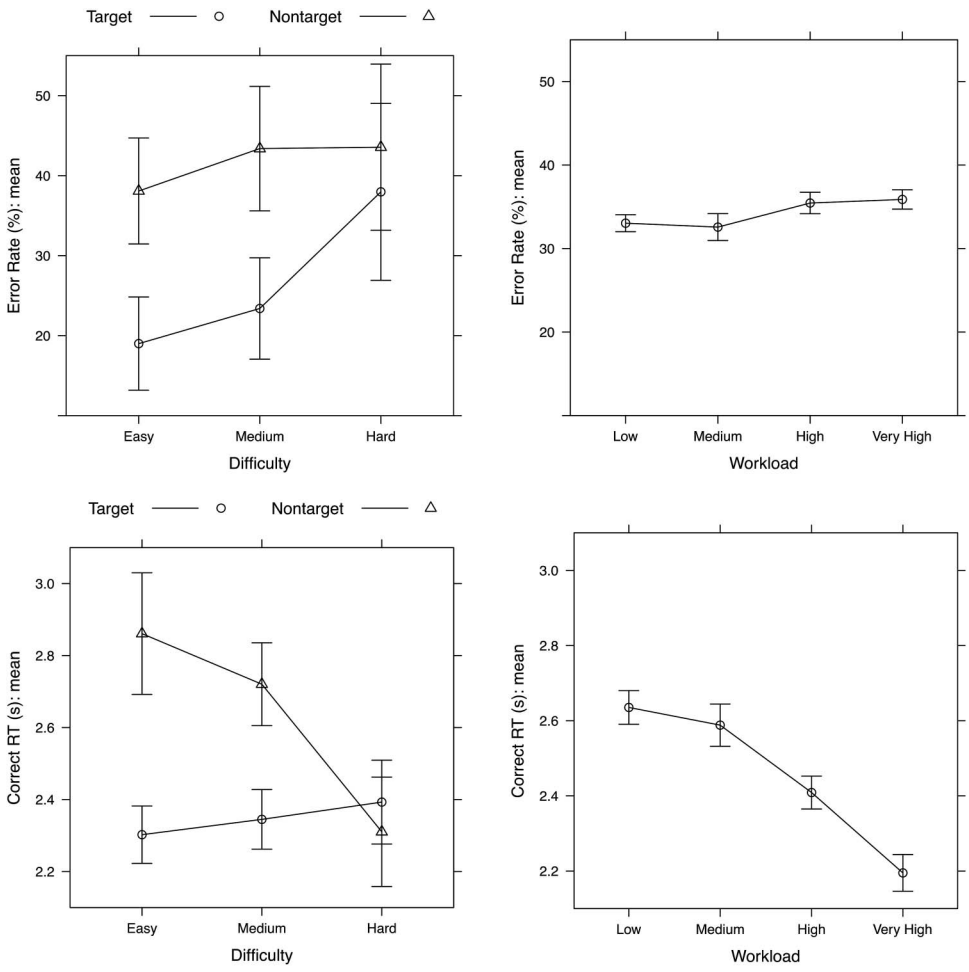

Figure 5. The interaction between difficulty and stimulus type (left column) and the main effect of workload (right column) in error rates (upper row) and average reaction time (RT) for correct responses (lower row). Error bars represent 95% confidence intervals calculated using Morey's (2008) within-subject method.

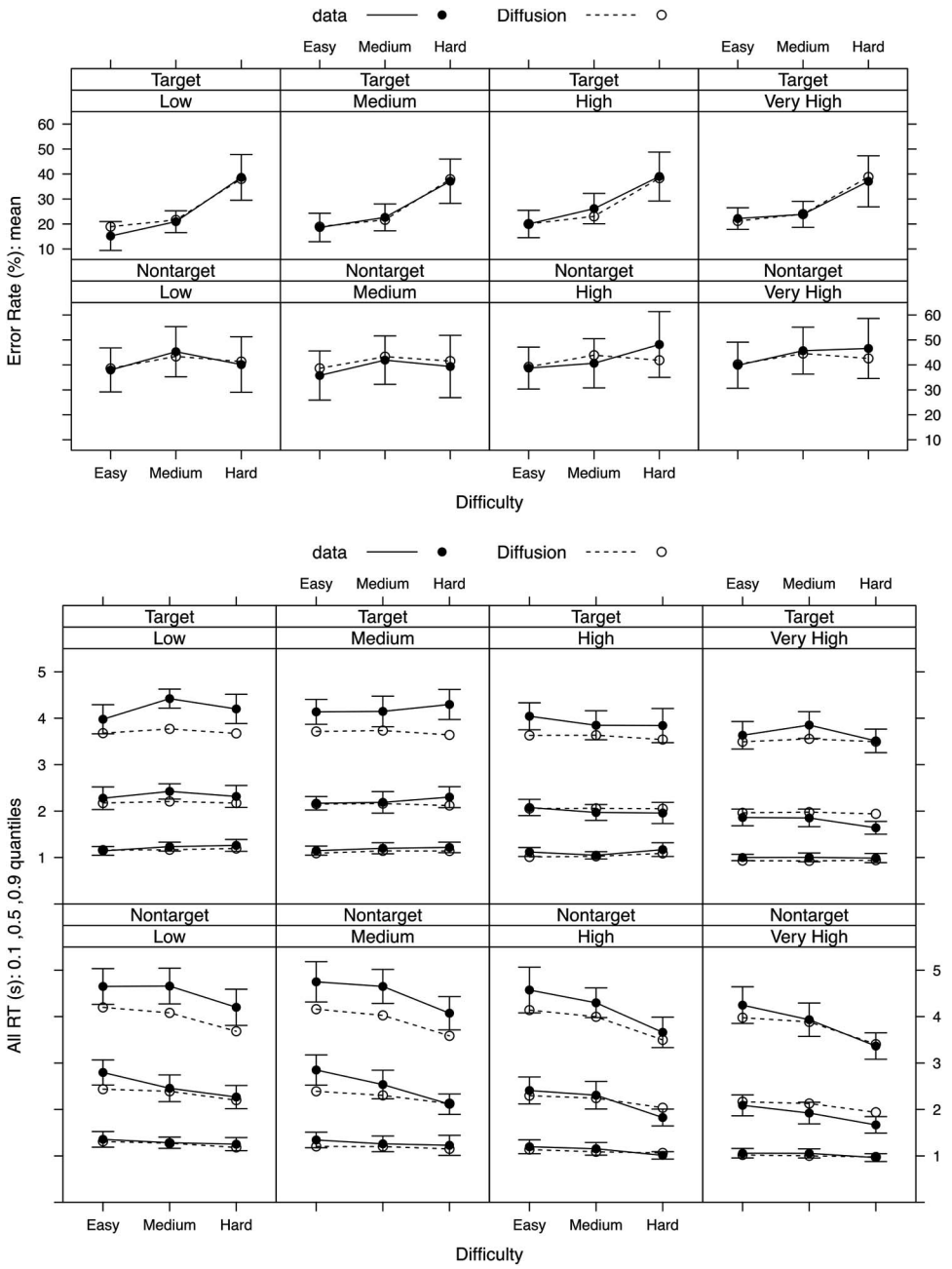

Figure 6. Fits of the AIC-selected diffusion model to error rates and RT distribution data with 95% confidence intervals. Panels from left to right represent increasing workload. AIC = Akaike Information Criterion; RT = reaction time.

can take account of response bias, but which are also sensitive to trade-offs between decision speed and accuracy. Our primary aim was to use these models to understand performance in a complex task similar to many applied decision scenarios: a UAV simulation task. We were interested in two effects that have large impacts on performance in applied setting, the sensory difficulty of the target discriminations (i.e., decision uncertainty) and time pressure or workload related to the number of simultaneously present potential targets. We fit two accumulate-to-threshold models to our data—the diffusion (Ratcliff & McKoon, 2008) and LBA (Brown & Heathcote, 2008). The purpose here was not to select which model was better, but rather to assess the degree to which any conclusions that we drew were sensitive to differences in their detailed assumptions.

Because ships could be present simultaneously, we were initially faced with the question of "cognitive architecture;" whether participants make decisions in series or in parallel. Although serial and parallel processing architectures can be very hard to differentiate in general (Townsend, 1990), the assumptions made by our models formed the basis for a clear inference in favor of serial processing. This inference was based on quantitative grounds

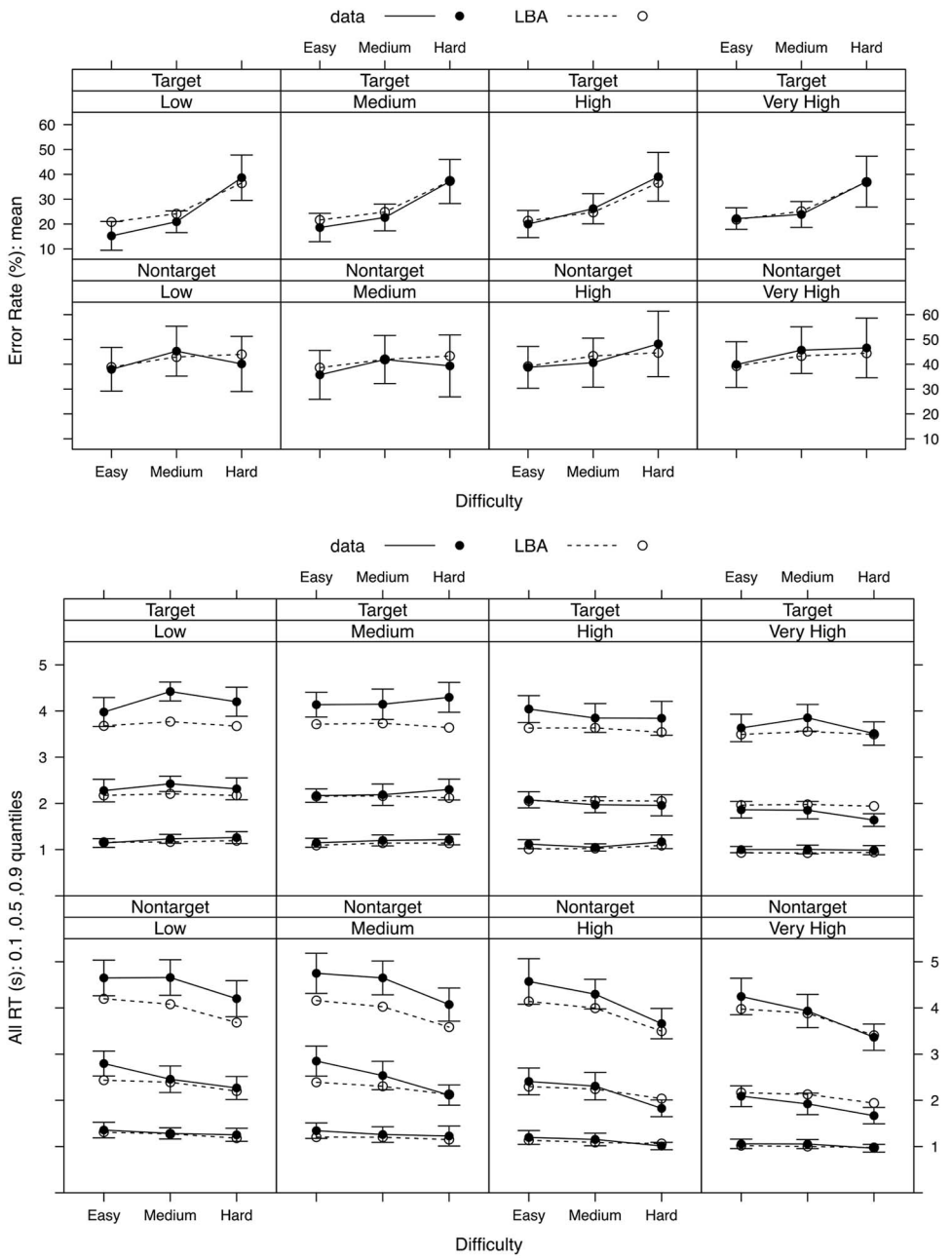

Figure 7. Fits of the AIC-selected LBA model to error rates and RT distribution data with 95% confidence intervals. Panels from left to right represent increasing workload. LBA = linear ballistic accumulation; AIC = Akaike Information Criterion; RT = reaction time.

(model fit) but also had a qualitative signature. The scoring method that corresponding to serial processing produced the skewed distribution of response time (RT) pervasively observed in simple tasks, whereas the method that corresponded to parallel processing did not. Although the validity of this inference is conditional on the veracity of the model assumptions, a large body of research supports them at a general level. Also, both models found support for serial processing, so this conclusion did not rely on their different detailed assumptions.

These findings are consistent with participants solving the problem of having only limited capacity to make choices by allocating all resources to one decision before moving on to the next. We note that this conclusion is not trivial as our participants actually sped up responses as more ships were present, with only a modest increase in error rates. A speed up in responding can be produced by an increase in cognitive resources as workload increased (so-called "super-capacity"), but our modeling clearly rejected this possibility, as we now discuss.

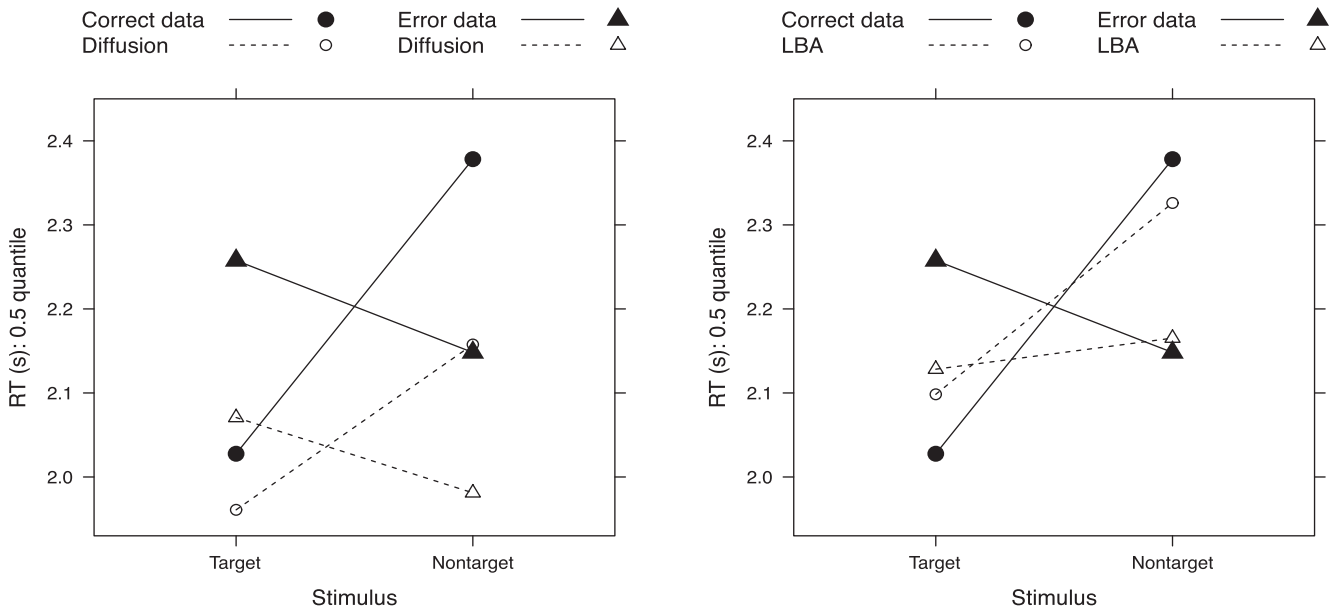

Figure 8. Fits of the AIC-selected diffusion (left) and LBA (right) models to the interaction between stimulus type and response accuracy in median RT. LBA = linear ballistic accumulation; AIC = Akaike Information Criterion; RT = reaction time.

Our analysis of behavior in the UAV task aimed to determine whether similar relationships between model parameters and experimental manipulations would emerge as in simple choice tasks. In particular, we examined two selective-influence hypotheses. First, we hypothesized that time pressure would selectively affect parameters concerned with the decision process itself, such as the model's threshold parameters, which determines the amount of accumulated evidence required to trigger a decision. Second, we hypothesized that difficulty would selectively affect parameters concerned with the inputs to the decision process, such as the mean and variability in the rate at which evidence is accumulated.

We reasoned that in our highly demanding task, where ships were only visible for a limited period of time, participants would make the strategic decision to set a lower threshold when more ships were present (i.e., when time pressure was higher) in order speed their decision processes, and so avoid failing to make a response at all. We operationalized difficulty by obscuring the ships with clouds of varying opacity, and reasoned that the quality of information supporting target versus nontarget discriminations would be affected, both in terms of the mean rate at which discriminating evidence could be extracted, and potentially in terms of variability in that rate from trial–trial, due to chance differences in the occlusion of critical features.

Our results clearly supported both selective influence hypotheses, and this support was consistent for both the diffusion and

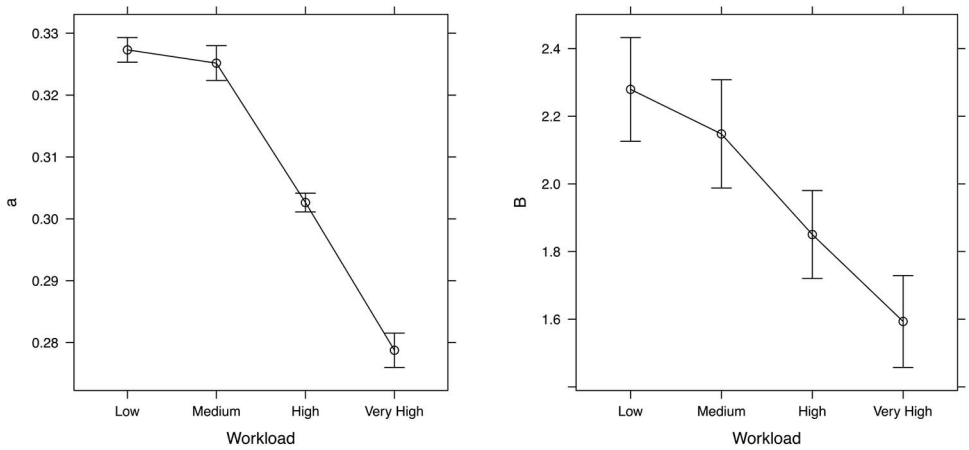

Figure 9. Threshold estimates for the diffusion (left) and the LBA (right) AIC selected models, with within-subject standard errors (Morey, 2008). LBA = linear ballistic accumulation; AIC = Akaike Information Criterion.

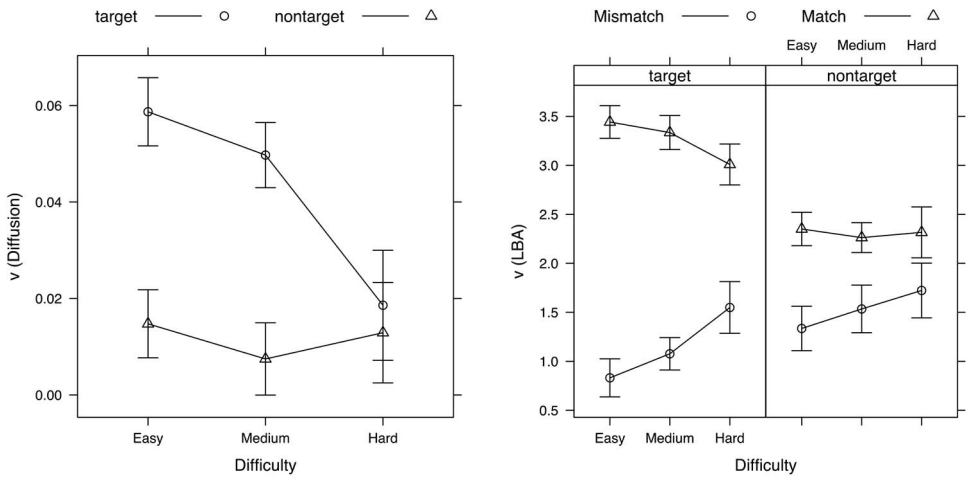

Figure 10. Mean rate estimates for the diffusion (left) and the LBA (right) AIC selected models, with within-subject standard errors (Morey, 2008). LBA = linear ballistic accumulation; AIC = Akaike Information Criterion.

LBA. In each, model variants where difficulty affected only mean rates and workload affected only the threshold were chosen as providing the best trade-off between goodness-of-fit and providing a parsimonious explanation.

With respect to thresholds, participants appeared to take advantage of the nonlinear relationship between speed and accuracy, which allowed them to substantially speed their choices as workload increased—and so minimize the number of missed responses—with only a minor penalty in terms of an increased error rate. Participants did not, however, appear to be able to increase their rate of evidence accumulation, or reduce its variability, as workload increased. That is, the strategy of increasing levels of attention or effort (i.e., a supercapacity effect) observed in simple redundant-target tasks (Eidels et al., 2010), did not seem to hold in our more complex and engaging task.

These findings are consistent with the possibility that time pressure effects on rate parameters in relatively easy simple tasks occur because participants were less engaged, so they had spare attentional capacity to deploy when required. It appears that our participants, in contrast, were working at full capacity even at the lowest level of time pressure, and as a result, had to adjust their threshold to accommodate increases in workload. However, there may well be circumstances under which people performing complex tasks do have spare capacity to deploy as workload levels increase, and are motivated to do so. For example, air traffic controllers can spend extended periods of time under low workload, if the aircraft under their jurisdiction are all flying at their cruising level and are on segregated routes. It is possible that air traffic controllers may respond to an increase in workload caused by new aircraft entering the sector, or aircraft requesting diversion around weather, by adjusting their rate (i.e., by working faster). They may only adjust their threshold as they approach their capacity limit (Loft, Bolland, Humphreys, & Neal, 2009; Neal et al., 2013).

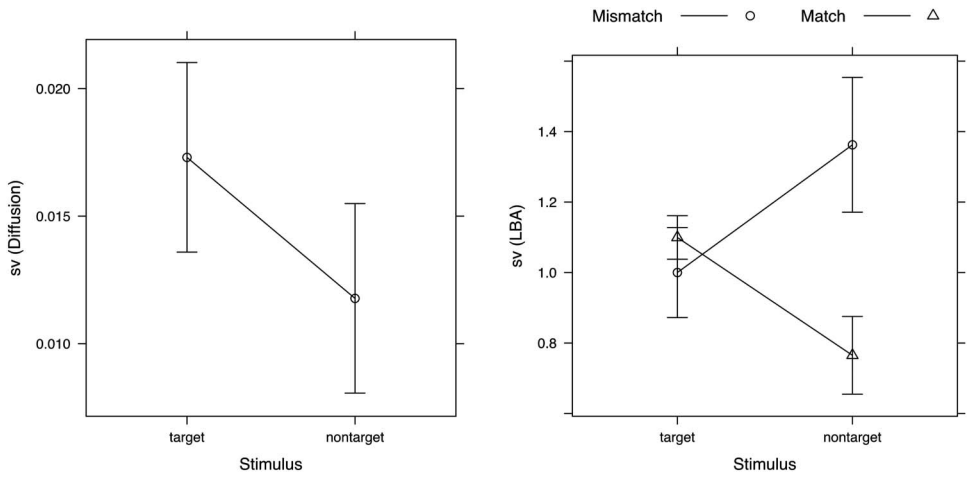

Figure 11. Rate standard deviation estimates for the diffusion (left) and the LBA (right) AIC selected models, with within-subject standard errors (Morey, 2008). LBA = linear ballistic accumulation; AIC = Akaike Information Criterion.

Table 4
Average Sizes of Error Rate Differences for the Differences Between Levels for Stimulus (Nontarget–Target) and Workload (Very High–Low) Factors, and for the Difficulty (Hard–Easy) Factor Separately for Target and Nontarget Stimuli

	Stimulus	Difficulty nontarget	Difficulty target	Workload
Data				
Average size	14.88	18.97	12.22	2.85
LBA				
Correlation	.99***	.98***	.94***	.05
Average size	14.36	15.44	10.28	.62
Diffusion				
Correlation	.99***	.98***	.90***	.24
Average size	14.71	18.53	10.6	1.58

Note. LBA = linear ballistic accumulation.
*** $p < .001$.

Difficulty was found to selectively influence the mean rate of evidence accumulation but not its variability (Loft, Sanderson, Neal, & Mooij, 2007). The latter finding suggests that there was not substantial variability from trial-to-trial trial in the degree to which the clouds obscured critical features. The fact that we found no effect of difficulty on thresholds, even though in principle it was possible and potentially advantageous to set lower threshold for more difficult discriminations, is consistent with threshold adjustment being an effortful and time consuming process. That is, if threshold adjustment must be completed before accumulation begins the associated delay might mitigate any advantage. Even if the threshold could be changed in parallel with evidence accumulation, it might be that the process of adjustment consumes limited attentional capacity and so decreases the rate of evidence accumulation, again limiting any advantage. However, it is possible that threshold adjustments may occur as a function of difficulty in other settings where the advantage to be gained is larger, and again this seems a fruitful topic for future research.

Model selection consistently indicated differences between targets and nontargets in trial-to-trial rate variability for both the diffusion and LBA. We had initially hypothesized that heterogeneity of our target and nontarget stimuli—which were defined in terms of a variety of features and feature conjunctions—could be accommodated by the model's rate-variability parameters. Although good model fits indicated this was the case, our findings indicated greater variability for targets than nontargets. To investigate further, we split targets into "easy" and "hard" types, based on whether they were defined by one or more features. When we repeated our model selection exercise we found that the stimulus effect on rate variability dropped out of the best model, which was also able to provide a good account of the large performance differences between hard and easy targets by allowing different mean rate parameters for hard and easy targets (see Appendix for details). However, this analysis was challenging, because the breakdown by target difficulty resulted in there being few trials in some design cells, particularly those associated with hard targets, because they occurred at half the rate of easy targets.

These findings illustrate a dilemma that will be faced in many applied setting with heterogeneous stimuli and other conditions that influence performance. An ever-finer breakdown by such influential factors offers the possibility of a more accurate and nuanced account in terms of parameter differences between conditions. Eventually, however, the number of trials available in each condition becomes insufficient to support model estimation. That is, inevitably there is a trade-off that requires some judgment by the researcher, both in terms of task design and model specification, with parameters that allow for variability within a condition providing some leeway. In the present case, for example, the trial-to-trial rate variability parameter seem to have functioned in this way, allowing the models to provide a reasonably accurate account of behavior and a sensible interpretation of its causes even when the large effect of target difficulty was neglected.

A related dilemma concerns practical limitation in applying the approach used here of fitting a very large number of model variants. Computational cost means that is not possible to explore every conceivable variant. For example, we did not test whether stimulus encoding time, and hence total nondecision time, differed depending on workload,[6] or whether response bias changes with workload. The good account of fast RTs (which are most sensitive to nondecision time effects), and of differences between responses to each stimulus (which are sensitive to response bias) indicates that such effects are less likely. However, it is also possible that any misfit would be masked by a trade-off in other parameter values, and we acknowledge that our conclusions must be tempered by these restrictions. Future research investigating, and careful consideration of these assumptions in other tasks, would seem prudent.

Although both models did provide a quite accurate account of behavior, particularly in terms of accuracy, there were also clear shortcomings. One of these was a consistent underestimation of slower responses, which is not seen, or at least is greatly attenuated, in simple tasks (e.g., Heathcote, Loft, & Remington, 2015; Heathcote & Love, 2012; Rae et al., 2014). That is, more slow responses were observed than predicted by the models. We speculated that our sequential method of scoring RT, in terms of the time since the last decision, might cause a spurious inflation of RT on some trials. In particular, this could occur where participants focus on a ship for some period of time, but it then disappears before they feel they have sufficient information to respond. The sequential scoring method would then result in a slow RT on the next trial, as it contains the time

[6] Because the LBA can be subject to tradeoffs between t_{er} and B effects we also investigated models where t_{er} was allowed to vary as a function of workload, but these models were not selected by AIC. However, parity with the diffusion model made it more straightforward to report results for the restricted LBA variants with only one estimated t_{er} value.

Table 5
Average Sizes of Mean Reaction Time Differences for the Differences Between Levels for the Workload (Very High–Low) Factor, for the Stimulus (Nontarget–Target) Factor Separately for Correct and Error Responses, and for the Difficulty (Hard–Easy) Factor Separately for Target and Nontarget Stimuli

	Stimulus correct	Stimulus error	Difficulty nontarget	Difficulty target	Workload
Data					
Average size	.39	−.16	−.48	.01	−.50
LBA					
Correlation	.84***	.67***	.78***	.82***	.48**
Average size	.26	.06	−.29	−.01	−.24
Diffusion					
Correlation	.88***	.43*	.53**	.50**	.74***
Average size	.16	−.03	−.17	.09	−.39

Note. LBA = linear ballistic accumulation.
* $p < .05$. ** $p < .01$. *** $p < .001$.

for partially processing the previous stimulus. Over trials variation in the amount of partial processing could also inflate variability in RT. Both effects could cause an exaggeration in the slow tail of RT distribution, and hence cause it to be underestimated by the models. The limitations of our model-based inferences regarding serial versus parallel processing could be addressed by controlling (e.g., by selectively revealing details of potential targets) or measuring (e.g., by eye-tracking) attention allocation strategies.

The navigation alarm sounding during the course of a choice could also cause distraction, and hence slow responding, as could many other factors extraneous to the choice task that will often be present in applied settings. Based on these considerations, it is possible that underestimation of slow RTs could be a pervasive phenomenon in many real-world scenarios. Two ways forward are afforded by a model-based approach. One is to explicitly account for distracting processes. However, estimating parameters related to distraction may be challenging unless distracting events are relatively common. Another approach is to use the fit of a standard model as a benchmark against which to measure the degree of distraction. At the very least, this sort of normative use standard models provides a way to assess the need for more elaborate modeling, and perhaps as an indicator or the need for control measures to mitigate distracting influences.

The models provided an excellent account of the average size of workload and difficulty effects on accuracy, and in particular the marked interaction between stimulus and difficulty, with the stimulus affect being larger on targets than nontargets. This account extended beyond the average of over participants to individual differences, with correlations between model predictions and participant performance mostly above .90 for all of the substantial accuracy effects. However, there were shortcomings in the account provided of corresponding RT effects, with the predictions being in the correct directions but generally only half to two thirds the observed size. Correlations with individual differences were also less in magnitude, between .50 and .80 for the substantial effects, but were still clearly significant.

Despite these shortcomings the models were able to provide a relatively simple account of what was on the surface seems a very complex pattern of differences both within and between RT and accuracy measures, particularly in relation to interactions with stimulus effects. Error rate increased only slightly with workload (3%), whereas RT decreased markedly, being over 0.5 s faster on average under very-high workload than low workload. Error rates for targets increased markedly with difficulty (19%), whereas the increase for nontargets was smaller (12%). In contrast, RT for nontargets decreased markedly with difficulty, by more than 0.5 as, whereas it was virtually unaffected for targets.

This complex pattern emerged from two main types of changes in model parameters: a decrease in thresholds as workload increased, and mean rates (or rate differences for LBA) that were greater for nontargets than targets. The latter rate difference was largest for the least occluded stimuli, but tended to disappear as occlusion increased. The threshold effect has a straightforward explanation as a strategic response to minimize nonresponses. This strategy was not entirely effective, as nonresponse rates still increased as workload increased, but the decrease in threshold would at least slow the increase in nonresponses. An increase in the nonresponse rate with difficulty was also observed, but there was no corresponding adjustment of threshold, suggesting there are limits to participants' abilities to respond to factors causing nonresponses.

The effect of stimulus type on mean rates suggests that there were objective differences in the discriminability of the targets and nontargets that we used in the UAV task. That these differences tended to disappear as occlusion increases is plausibly a natural consequence of less information being available, as in the limit of complete occlusion there can be no differences. As previously discussed, a more nuanced explanation of rate effects, including the effect of stimulus on rate variability, is provided by an analysis accounting for differences among types of targets reported in the Appendix.

The models also provided a simple explanation—in terms of threshold-meditated response bias in favor of targets—for the substantial interaction observed between response accuracy and stimulus type: Correct responses to targets were faster than error responses by 0.4 s, whereas correct responses to nontargets were slower than errors responses by 0.16 s. A lower target threshold leads to faster target responses, as it takes less time to accrue the required amount of evidence. This is true both when a target response is correct (i.e., when made to a target stimulus) and when it is incorrect (i.e., when made to a nontarget stimulus).

In conclusion, we believe that, even though it underestimated some RT effects, the model-based analysis was valuable in providing a psychologically meaningful distillation of the complex behavior observed in our complex task. More broadly we believe the present results encourage the wider application of accumulate-to-threshold models in applied settings. In the longer term, such applications will likely identify shortcomings that will require elaboration of the

standard models, but this process of model development is also likely to be illuminating about important psychological processes. In the shorter term, and particularly in light of how well they handled the relatively complex and challenging nature of our UAV task and the behavior it engendered, applications of standard model appear to have great promise for a better understanding of how applied decision-makers adapt to dynamic and uncertain environments.

Safety-critical jobs, such as air traffic control and maritime or UAV surveillance, require operators to make decisions in complex and uncertain environments. Understanding the psychological processes underlying decision-making in these demanding environments is essential for improving the safety and efficiency of these work systems. Our study applied standard accumulate-to-threshold models to address this issue, revealing that these standard models can capture people's strategic responses to changes in time pressure and uncertainty.

Extending the application of these models to more complex tasks such as ours shows that these models can lend themselves to tasks more representative of applied settings, which have typically been modeled using SDT. A shortcoming of SDT is unable to account for response time—an outcome that is intricately related to accuracy. Using accumulate-to-threshold models to understand decision-maker's performance could serve as a more sophisticated and generalizable measurement model to identify conditions where existing work systems require redesign, or evaluate the efficacy of human-technology interfaces.

References

Arnold, N. R., Bröder, A., & Bayen, U. J. (2014). Empirical validation of the diffusion model for recognition memory and a comparison of parameter-estimation methods. *Psychological Research.* Advance online publication.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7*. Retrieved from http://CRAN.R-project.org/package=lme4

Brown, S., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review, 112,* 117–128. http://dx.doi.org/10.1037/0033-295X.112.1.117

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

Brown, S. D., Marley, A. A. J., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review, 115,* 396–425. http://dx.doi.org/10.1037/0033-295X.115.2.396

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33,* 261–304. http://dx.doi.org/10.1177/0049124104268644

Diederich, A. (1997). Dynamic stochastic models for decision making with time constraints. *Journal of Mathematical Psychology, 41,* 260–274. http://dx.doi.org/10.1006/jmps.1997.1167

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

Donkin, C., Brown, S. D., & Heathcote, A. (2011). Drawing conclusions from choice response time models: A tutorial using the linear ballistic accumulator. *Journal of Mathematical Psychology, 55,* 140–151. http://dx.doi.org/10.1016/j.jmp.2010.10.001

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

Fific, M., Little, D. R., & Nosofsky, R. M. (2010). Logical-rule models of classification response times: A synthesis of mental-architecture, random-walk, and decision-bound approaches. *Psychological Review, 117,* 309–348. http://dx.doi.org/10.1037/a0018526

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Hawkins, G. E., Marley, A. A. J., Heathcote, A., Flynn, T. N., Louviere, J. J., & Brown, S. D. (2014). Integrating cognitive process and descriptive models of attitudes and preferences. *Cognitive Science, 38,* 701–735. http://dx.doi.org/10.1111/cogs.12094

Heathcote, A., & Hayes, B. (2012). Diffusion versus linear ballistic accumulation: Different models for response time with different conclusions about psychological mechanisms? *Canadian Journal of Experimental Psychology, 66,* 125–136. http://dx.doi.org/10.1037/a0028189

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

Heathcote, A., & Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in Psychology, 3,* 292.

Heathcote, A., Wagenmakers, E. J., & Brown, S. D. (2014). The falsifiability of actual decision-making models. *Psychological Review, 121,* 676–678. http://dx.doi.org/10.1037/a0037771

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: A theory of the relationship between individual differences and information processing. *Psychological Review, 91,* 153–184. http://dx.doi.org/10.1037/0033-295X.91.2.153

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

Karayanidis, F., Mansfield, E. L., Galloway, K. L., Smith, J. L., Provost, A., & Heathcote, A. (2009). Anticipatory reconfiguration elicited by fully and partially informative cues that validly predict a switch in task. *Cognitive, Affective & Behavioral Neuroscience, 9,* 202–215. http://dx.doi.org/10.3758/CABN.9.2.202

Laming, D. R. J. (1968). *Information theory of choice-reaction times*. New York, NY: Academic Press.

Little, D. R., Nosofsky, R. M., & Denton, S. E. (2011). Response-time tests of logical-rule models of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1–27. http://dx.doi.org/10.1037/a0021330

Little, D. R., Nosofsky, R. M., Donkin, C., & Denton, S. E. (2013). Logical rules and the classification of integral-dimension stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 801–820. http://dx.doi.org/10.1037/a0029667

Loft, S., Bolland, S., Humphreys, M., & Neal, A. (2009). A theory and model of conflict detection in air traffic control: Incorporating environmental constraints. *Journal of Experimental Psychology: Applied, 15,* 106–124. http://dx.doi.org/10.1037/a0016118

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,* 376–399. http://dx.doi.org/10.1518/001872007X197017

Luce, R. D. (1986). *Response times*. New York, NY: Oxford University Press.

McCarley, J. S., & Wickens, C. D. (2005). *Human factors implications of UAVs in the National Airspace. Tech. Rep. No. AHFD-05–05/FAA-05–01*. Atlantic City, NJ: Federal Aviation Administration.

Miller, J. (1978). Multidimensional same—Different judgments: Evidence against independent comparisons of dimensions. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 411–422. http://dx.doi.org/10.1037/0096-1523.4.3.411

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology, 14,* 247–279. http://dx.doi.org/10.1016/0010-0285(82)90010-X

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4,* 61–64.

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,* 287–305. http://dx.doi.org/10.1177/0018720813491283

Neal, A., & Kwantes, P. J. (2009). An evidence accumulation model for conflict detection performance in a simulated air traffic control task. *Human Factors, 51,* 164–180. http://dx.doi.org/10.1177/0018720809335071

Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1226–1243. http://dx.doi.org/10.1037/a0036801

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

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review, 9,* 278–291. http://dx.doi.org/10.3758/BF03196283

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20,* 873–922. http://dx.doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9,* 347–356. http://dx.doi.org/10.1111/1467-9280.00067

Schmiedek, F., Oberauer, K., Wilhelm, O., Süss, H. M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General, 136,* 414–429. http://dx.doi.org/10.1037/0096-3445.136.3.414

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology, 64,* 1–34. http://dx.doi.org/10.1016/j.cogpsych.2011.10.002

Stone, M. (1960). Models for choice-reaction time. *Psychometrika, 25,* 251–260. http://dx.doi.org/10.1007/BF02289729

Townsend, J. T. (1990). Serial vs. parallel processing: Sometimes they look like Tweedledum and Tweedledee but they can (and should) be distinguished. *Psychological Science, 1,* 46–54. http://dx.doi.org/10.1111/j.1467-9280.1990.tb00067.x

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology, 39,* 321–359. http://dx.doi.org/10.1006/jmps.1995.1033

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review, 121,* 179–205. http://dx.doi.org/10.1037/a0036137

Trueblood, J. S., Brown, S. D., Heathcote, A., & Busemeyer, J. R. (2013). Not just for consumers: Context effects are fundamental to decision making. *Psychological Science, 24,* 901–908. http://dx.doi.org/10.1177/0956797612464241

Vuckovic, A., Kwantes, P. J., Humphreys, M., & Neal, A. (2014). A sequential sampling account of response bias and speed-accuracy tradeoffs in a conflict detection task. *Journal of Experimental Psychology: Applied, 20,* 55–68. http://dx.doi.org/10.1037/xap000007

Vuckovic, A., Kwantes, P. J., & Neal, A. (2013). Adaptive decision making in a dynamic environment: A test of a sequential sampling model of relative judgment. *Journal of Experimental Psychology: Applied, 19,* 266–284. http://dx.doi.org/10.1037/a0034384

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language, 58,* 140–159. http://dx.doi.org/10.1016/j.jml.2007.04.006

Whitson, L. R., Karayanidis, F., Fulham, R., Provost, A., Michie, P. T., Heathcote, A., & Hsieh, S. (2014). Reactive control processes contributing to residual switch cost and mixing cost across the adult lifespan. *Frontiers in Psychology, 5,* 383.

Appendix

Easy Versus Hard Targets

One intriguing finding from model selection was an effect of stimulus on rate variability. Analysis of the parameters from the AIC-selected model revealed that target rate variability was significantly larger than nontarget rate variability in the LBA, and that the same trend was present for the diffusion model. A potential explanation for this finding is that the stimuli classified as targets were more heterogeneous in the classification-relevant evidence they afforded than nontargets. When we broke down the data by target type a striking pattern was revealed; accurate and fast performance for the two-thirds of targets that were defined only in terms of its hull color (green), but very error prone and much slower performance for the remaining one third of targets with a much more complicated definition, a conjunction of the remaining two hull colors (black or yellow) and having no cranes or more than one crane.

Here we report the results of exploratory analyses breaking down the two-level stimulus factor (nontargets vs. targets) into three levels: nontargets, easy targets (i.e., green-hulled ships), and hard targets (i.e., the remaining target types). This further breakdown, combined with the relative rarity of hard targets and missing data due to nonresponding, meant that we had to drop one participant from this analysis, as they had little or no data in several cells (no responses in two cells and only one in three others). Even for some of the participants retained there were not many observations for hard targets (on average nine in each of the 12 combinations of workload and difficult), so the following results have to be viewed with some caution.

(*Appendix continues*)

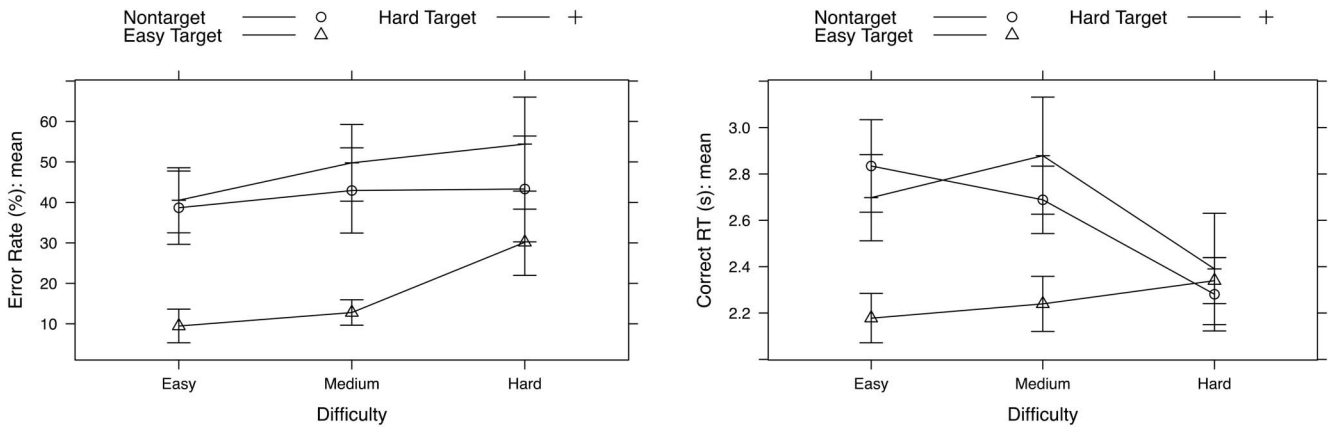

Figure A1. The interaction between difficulty and stimulus type (broken down by hard and easy targets) on: error rates (left) and average reaction time (RT) for correct responses (right). Error bars represent 95% confidence intervals calculated using Morey's (2008) within-subject method.

Results

Nonresponse rates were greater for nontargets and targets compared with easy targets (10% 10%, 7%), $\chi^2(2) = 25.04$, $p < .001$. Consistent with original results, nonresponse rates increased as difficulty increased (6%, 9%, 12%), $\chi^2(2) = 45.69$, $p < .001$, and as workload increased (7%, 8%, 9%, 10%), $\chi^2(3) = 123.44$, $p < .001$.

Error Rates and Mean Correct RT

The same pattern of effects on mean correct RT and error rates shown in Table 3 occurred with the easy/hard target breakdown (we did not analyze the other measures in Table 3 due to low numbers of observations). Mean correct RT decreased as workload increased, $\chi^2(3) = 104.08$, $p < .001$ (2.77s, 2.65s, 2.47s and 2.21s), and accuracy decreased, $\chi^2(2) = 12.53$, $p = .006$ (65%, 65%, 63%, and 63%). The main effect of stimulus was highly significant for both correct mean RT, $\chi^2(2) = 79.95$, $p < .001$, and error rates, $\chi^2(2) = 1259.68$, $p < .001$. In both cases this was the overall level of performance for hard targets was much less than that for easy targets and close to that of nontargets (mean correct RT: 2.65 s, 2.27 s, and 2.66 s, error rates: 48%, 17%, and 42%, respectively).

There was a significant interaction between stimulus (nontarget, easy target, and hard target) and difficulty, both for mean correct RT, $\chi^2(4) = 46.07$, $p < .001$, and error rates, $\chi^2(4) = 211.88$, $p < .001$. The reasons were the same as for the analysis aggregating over target type; for mean correct RT the largest effect of difficulty was on nontargets, whereas for accuracy its largest effect was on targets. However, there were some qualifications on the similarity of the interactions, as shown in Figure A1. For mean correct RT responses to hard targets in low difficulty condition were faster for easy targets, but the opposite was true in the medium difficulty condition. For error rates, the decrease in accuracy as difficulty increases was steeper for easy than hard targets.

Model Fitting

We fit the diffusion and LBA models to the data broken down by hard and easy targets in the same way as before, except the top model was the AIC selected model from previous analyses (more complex models led to unstable estimates due to low numbers of observations). Our interest focused on whether the breakdown by target difficulty removed its effect on the rate standard deviation (s_v) parameter. As Tables A1 and A2 show, this was indeed the case

Table A1
Diffusion Fits Over Participants

Variant	*a*	*v*	s_v	*z*	*p*	D	AIC
Top	WL	S,D	S	–	20	74061	75181
AIC	WL	S,D	—	—	18	74116	75124

Note. D = deviance; AIC = Akaike Information Criterion; *p* = number of parameters per participant. Model factors are WL = workload; D = difficulty; S = stimulus broken down by target difficulty.

(*Appendix continues*)

Table A2
Linear Ballistic Accumulation Fits Summed Over Participants

Variant	*B*	*v*	s_v	*p*	D	AIC
Top	R,WL	S,D,M	S	33	72520	74368
AIC	R,WL	S,D,M	—	29	72708	74332

Note. D = deviance; AIC = Akaike Information Criterion; *p* = number of parameters per participant. Model factors are WL = workload; D = difficulty; S = stimulus broken down by target difficulty. Accumulator factors are M = match; R = response.

for both the diffusion and LBA models. As before the LBA model was also selected over the diffusion model.

Figure A2 plots the fit of the AIC selected variants of the diffusion and LBA models. The fits aggregate over workload and omit the nontarget data in order to focus on the model's ability to fit the very large differences between the two types of targets. As before, the fit to accuracy is very good, but for RT the slow tail of RT distribution is systematically underestimated.

Figures A3 and A4 plot threshold and mean rate estimates for the AIC selected model. The pattern of effect on thresholds is the same as in Figure 9. For the diffusion the effect of workload on threshold remained significant, $\chi^2(3) = 281.01$, $p < .001$, but for the LBA it became nonsignificant, although it was numerically similar, $\chi^2(3) = 4.04$, $p = .257$. For diffusion's mean rates the main effects of difficulty, $\chi^2(2) = 9.97$, $p = .006$, and stimulus, $\chi^2(2) = 77.43$, $p < .001$, were again significant. As Figure A4 shows, mean rates were much higher for easy than hard targets, with the latter being similar to non-targets. The interaction between these two factors

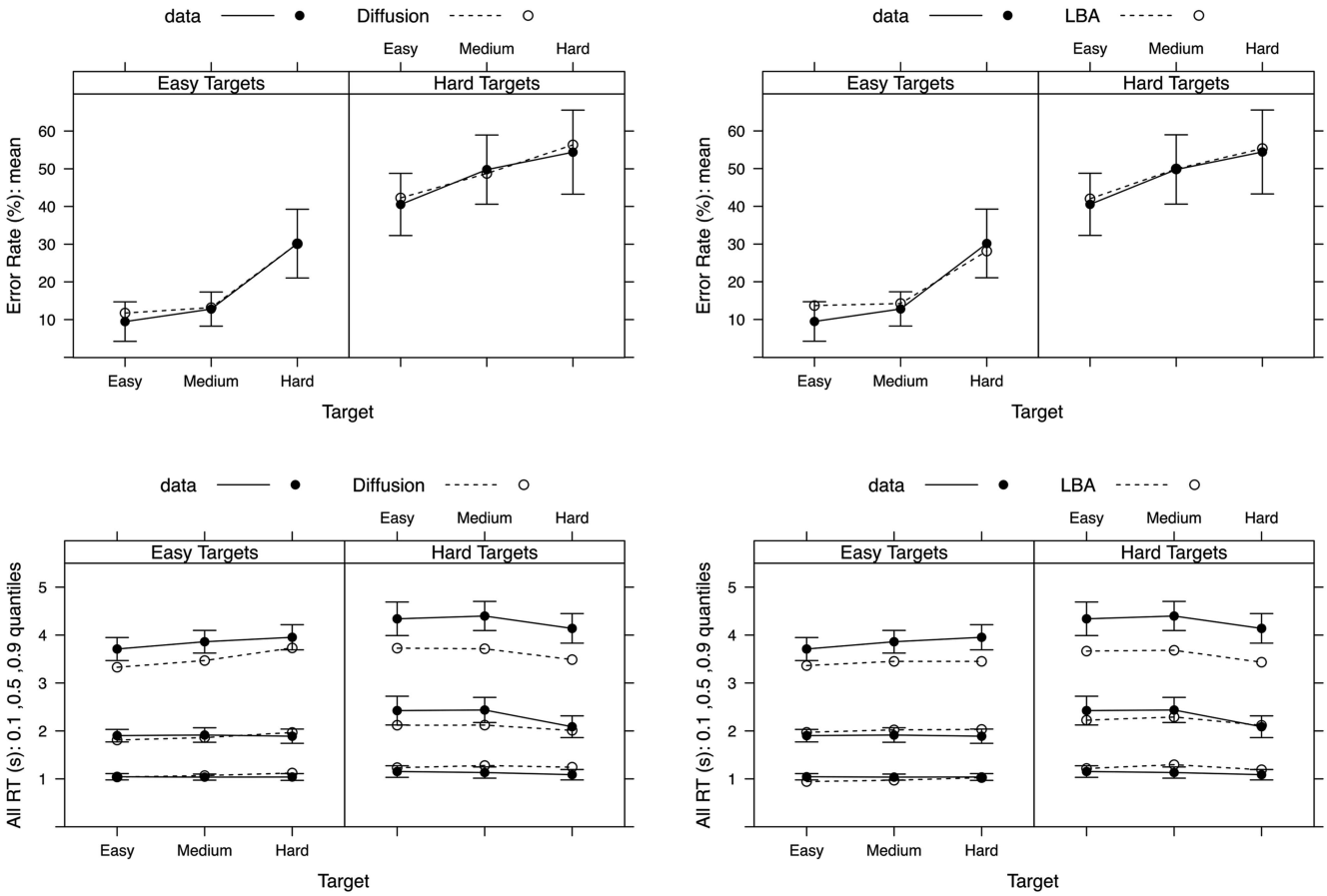

Figure A2. Fits of the AIC-selected diffusion and LBA models to error rates and RT distribution data for hard and easy targets, with 95% confidence intervals. LBA = linear ballistic accumulation; AIC = Akaike Information Criterion; RT = reaction time.

(*Appendix continues*)

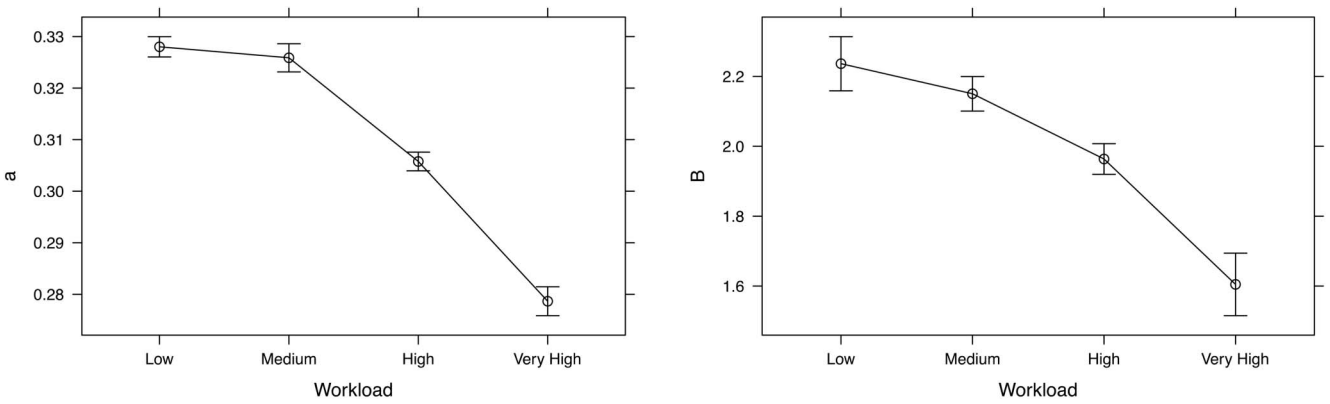

Figure A3. Threshold estimates for the diffusion (left) and LBA (right) AIC selected models for the data broken down by target type, with within-subject standard errors (Morey, 2008). LBA = linear ballistic accumulation; AIC = Akaike Information Criterion.

became only marginally significant, $\chi^2(4) = 7.83$, $p = .098$. For the LBA the same pattern of mean rate effects was found as before, with main effects for stimulus, $\chi^2(2) = 23.67$, $p < .001$, and match, $\chi^2(2) = 129.22$, $p < .001$, and interactions of match with stimulus, $\chi^2(2) = 156.04$, $p < .001$, and difficulty, $\chi^2(2) = 25.04$, $p < .001$. As Figure A4 shows, the pattern of effects was very similar to Figure 10, except that the match effect for hard targets was small, reflecting low levels of accuracy.

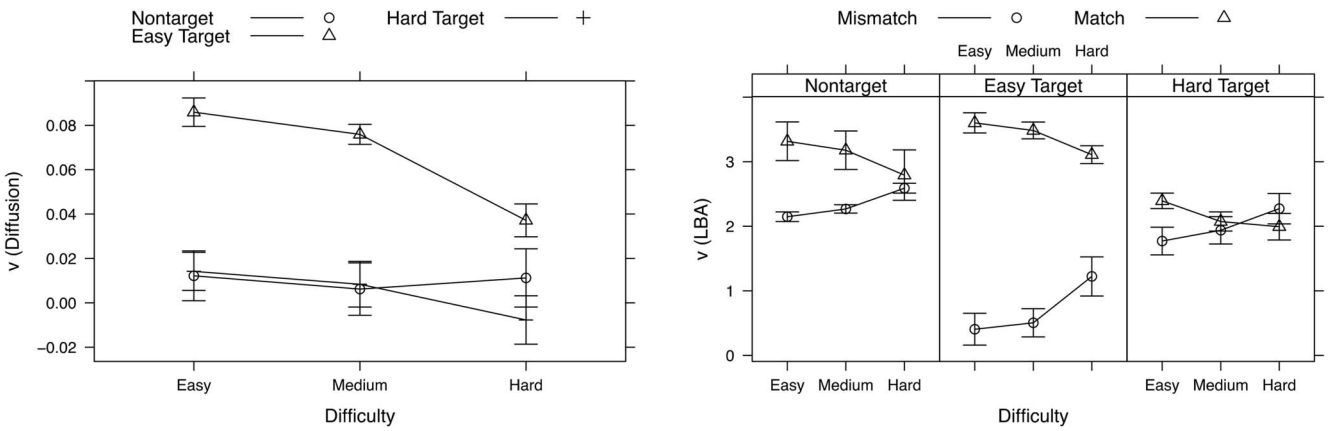

Figure A4. Mean rate estimates for the diffusion (left) and LBA (right) AIC selected models for the data broken down by target type, with within-subject standard errors (Morey, 2008). LBA = linear ballistic accumulation; AIC = Akaike Information Criterion.

Received August 20, 2015
Revision received November 7, 2015
Accepted November 10, 2015 ■