Less is More: How Individual Differences in Multi-tasking Ability Interact with Low-Degree Automation to Determine Task Performance, Cognitive Workload, and Situation Awareness

Jayden Greenwell-Barnden

Graduate Diploma of Science (Honours), Bachelor of Arts

This thesis is presented for the degree of Doctor of Philosophy

of The University of Western Australia

School of Psychological Science

2022

Thesis Declaration

I, Jayden Nicholas Greenwell-Barnden, certify that:

This thesis has been substantially accomplished during enrolment in this degree.

This thesis does not contain material which has been submitted for the award of any other degree or diploma in my name, in any university or other tertiary institution.

In the future, no part of this thesis will be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of The University of Western Australia and where applicable, any partner institution responsible for the joint-award of this degree.

This thesis does not contain any material previously published or written by another person, except where due reference has been made in the text and, where relevant, in the Authorship Declaration that follows.

This thesis does not violate or infringe any copyright, trademark, patent, or other rights whatsoever of any person.

The research involving human data reported in this thesis was assessed and approved by The University of Western Australia Human Research Ethics Committee. Approval #: RA/4/1/9020.

This research was supported through an Australian Government Research Training Program Scholarship and the Australian Army and Defence Science Partnerships agreement of the Defence Science and Technology Group (ID 7120), as part of the Human Performance Research network (HPRnet).

Signature: 

Date: 28/10/2022

Abstract

This thesis aims to contribute to our understanding of human-automation teaming; namely 'why does automation benefit task performance?' and 'do some people benefit from automation differently to others?' Traditional answers to these questions involve investigating the nature of the automation (i.e., what type of automation is best for a given task?) However, such perspectives do not account for a critical component in the system – the human operator. Humans have inherent cognitive limits in their ability to perform multiple tasks, which is ubiquitous in the modern workplace. These limits include our inability to divide attention, interference from competing tasks, and costs of switching between tasks. Due to the imperfect nature of almost all automation, which may not adapt to dynamic or complex situations or may malfunction, it is often not possible or practical to automate all aspects of a task, particularly when the cost of error is high. For this reason, the combination of humans and automation is essential in many workplaces, leading to the critical question of how to get the best out of the component parts of a system.

I will examine recent and emerging perspectives on this question, combining human factors and cognitive psychology approaches to determine if (and if so, how) cognitive abilities shape outcomes when teamed with automation. Automation typically benefits task performance by freeing-up cognitive resources of the human operator and may impact related factors, including workload and situation awareness. Thus, in theory, automation should have a proportionally greater influence on outcomes for operators who have fewer cognitive resources than for operators who have more cognitive resources. To date, however, this assumption has not been empirically tested.

The present experiments will test a novel framework that investigates whether profiling an operator's task-relevant cognitive ability (specifically multi-tasking) can predict their performance (including speed and accuracy) when using automation that assists that cognitive ability. Operator multi-tasking ability was assessed by combining scores from several short cognitive tasks that tapped different aspects of multi-tasking with a latent factor methodology. A simulated air traffic control task was chosen to represent the demands of a modern multi-tasking workplace where automation is deployed to assist human operators. Performance outcomes assessed include workload, situation awareness as well as task accuracy and response times. A common taxonomy combining the level of control exerted by automation and the stage of human information processing augmented by the automation is called the 'degree of automation'. Until now, prior literature has only looked for variations in outcomes *across* increasing degrees of automation. By contrast this thesis explores variations of performance outcomes *within* a single degree of automation.

This thesis is presented as a series of experiments prepared for publication. **Chapter 1**, the General Introduction, provides a synthesis of the relevant literature and summary of the research, including the overarching rationale for this thesis. **Chapter 2** is the first experimental chapter which established the theoretical framework and methodological approach, asking the question 'does multi-tasking interact with the presence or absence of a low-degree of automation to predict relevant outcomes?' It found poorer multi-taskers benefitted more from automation than better multi-taskers as they showed a greater improvement in their performance with automation compared to their manual performance. Better multi-taskers performed better, had better situation awareness, and reduced objective workload. **Chapter 3** replicated and extended the findings of Chapter 2 by applying the same automation to a distinct but related task in the air traffic simulation and introduced new and improved task performance measures. Together, the experimental **Chapters 2** and **3** examined how individual cognitive ability in multi-tasking predicts automation's benefits to task performance (i.e., do some people benefit from automation more than others?) and impact workload and situation awareness. **Chapter 4** examined if individual cognitive abilities in multi-tasking also predict the costs of automation error (i.e., do some people perform worse than others when automation makes an error?). It found that performance differed between manual, reliable automation, and automation error trials, and this was predicted by multi-tasking ability. Finally**, Chapter 5**, the General Discussion, summarised the findings of the experiments and concludes with broader theoretical and practical implications of this research.

Acknowledgements

There are many people to who I would like to express my gratitude and appreciation for making my PhD journey possible. Firstly, my coordinating supervisor Troy Visser (Assoc. Professor). You have been the cornerstone of my PhD, as having decided that while I could learn to enjoy any topic I studied, the working relationship with my supervisor would be the basis for a successful journey. I chose you for the confidence you inspired, your kindness, unbridled enthusiasm, and great knowledge. I saw these qualities from the lecture hall seats in undergraduate classes with you, and you continue to embody them all these years later. Your dedication to your role as supervisor was commendable. I will remember the banter, the Star Wars décor, and the critical importance of devoting time to small talk.

I also extend my thanks to my secondary supervisor Shayne Loft (Professor). Your expertise in human factors has been instrumental. Also, my heartfelt thanks go to Dr Vanessa Bowden, who has been an immense support from the very start. Thank you for your patience throughout. I would also like to thank Dr Angela Bender, who was integral to setting up my research in the early years, and for her excellent feedback on drafts. Also, thanks to Dr Susannah Whitney from the Defence Science and Technology Group for her encouragement and feedback. Special mentions to Dr Michael English for his help developing the tasks. Over the last four years, you have been a great help with your wealth of technical expertise and lending a sympathetic ear to my unfiltered thoughts. Thank you for patiently answering all my career questions and my 'what if…' thoughts. Thanks to my panel members, David Badcock (Professor) and Dr Serena Wee, for your feedback and encouragement in the early days. My gratitude also to all the many undergraduate students who have participated in my studies over the years, particularly under the difficult testing conditions during the 2020/21 Covid pandemic.

Importantly, I could not have completed the last four years without the love and support of my family and friends. Special mentions to my friends who were there for piloting (endless piloting) and participating in my studies: Kathleen, Daisy, Derek, and my office mate Chloe. Special mention of my best mates Brenton, Ryan, and Alex, who have listened, encouraged, and lent a hand at every turn. A most heartfelt thanks to my dear friend Cayla. Finally, thank you to my Mum. Your support throughout my academic journey have been constant, unconditional, and to me, immeasurable. I could not have achieved this without you. This research was supported through an Australian Government Research Training Program Scholarship and the Australian Army and Defence Science Partnerships agreement of the Defence Science and Technology Group (ID 7120), as part of the Human Performance Research network (HPRnet).

Table of Contents

List of Abbreviations

The following table describes the significant abbreviations and acronyms used throughout this thesis. The page on which each one is defined or first used is also given. Standard statistical acronyms are not in this list.

Abbreviation	Meaning	Page
AB	Attentional Blink	8
ATC	Air Traffic Control	12
DOA	Degree of Automation	17
DST	Defence Science Technology Group	
EFA	Exploratory Factor Analysis	69
LOA	Level of Automation	15
MOT	Multiple Object Tracking	163
MRT	Multiple Response Theory	9
NASA-TLX	NASA Task Load Index	22
OOTL	Out of the Loop	36
RSPM	Raven's Progressive Matrices	12
SA	Situation Awareness	4
SAGAT	Situation Awareness Global Assessment Technique	32
SART	Situation Awareness Rating Technique	32
SPAM	Situation Present Assessment Method	32
SWAT	Subjective Workload Assessment Technique	22
VST	Visuo-Spatial Tracking	92
WP	Workload Profile	32

Presentations Arising from this Thesis

Greenwell-Barnden, J., Bender, A., Loft, S., Whitney, S, & Visser, T., (November, 2019). *One size fits one: a new approach to human automation interaction*. Presented at the Defence Human Sciences Symposium, Canberra, Australian Capital Territory. Received best student presentation award.

Greenwell-Barnden, J., Bender, A., Loft, S., Whitney, S., & Visser, T. (November, 2020). *It's a Match! Improving Human-Automation Teaming by matching operator abilities to automation in the workplace*. Presented at the University of Western Australia School of Psychological Science Conference, Perth, Western Australia.

A Note About the Format of the Thesis

This thesis is presented as a series of papers prepared for publication, (called a 'thesis by publication' in Australia). This thesis begins with a General Introduction (**Chapter 1**) which reviews and synthesises the relevant literature in human factors and cognitive psychology domains, as this thesis is situated at the intersection of these fields, before presenting the overarching rationale for this thesis and summarising the general approach taken in each subsequent chapter. Three empirical studies are presented in chronological order of when they were conducted. **Chapter 2**, **Chapter 3,** and **Chapter 4** are being prepared for publication. As each experiment has been prepared as a standalone manuscript, there may be some overlap in the literature reviewed in the introduction/ discussion and the General Introduction/ General discussion of the thesis. Additionally, as the same tasks were used in each chapter (with some variation for experimental manipulation) there will be overlap in the methods section of each of the three experiments presented. This thesis concludes with the General Discussion (**Chapter 5**) which summarises the findings of the preceding studies and draws conclusions about their theoretical and practical implications considering the literature reviewed.

CHAPTER 1:

General Introduction

Chapter Overview

The first aim of this thesis is to investigate whether profiling individual differences in a task-relevant cognitive ability (e.g., multi-tasking) predicts performance outcomes. The second aim is to investigate whether using automation designed to augment cognitive ability results in better outcomes (i.e., improved task performance, maintained situation awareness and reduced workload) than without automation. Finally, the primary aim of this thesis is to investigate whether individual differences in multi-tasking ability vary automation's outcomes. These aims are described in more detail at the end of this chapter. Previous literature that will be reviewed in this chapter illustrate how differences in automated systems can vary human's performance, workload, and situation awareness outcomes. However, such studies have not adequately investigated the interaction between humans and automation at an individual level. Investigating how individual differences in cognition may interact with automation may demonstrate that the benefits and costs associated with automation are experienced differently by different people. Understanding this and accounting for individual cognitive ability may help design more efficient human-automation teaming systems in the future.

The first section of this chapter introduces multi-tasking, a cognitive ability critical to complex task performance in the modern workplace and explains why multi-tasking is inherently difficult due to limitations in its constituent cognitive processes. Explanatory models of these processes will then be reviewed. Multi-tasking is associated with individual differences across the population. Methods of measuring multi-tasking will be examined in detail, as this thesis uses several methods to operationalise and test multi-tasking ability.

Multi-tasking is directly aided by automation in many workplace environments to augment task performance. The second section will provide a detailed overview of human-automation teaming literature, first, by defining automation and a brief history of its conceptualisation with a focus on the prevailing 'degree of automation' taxonomy. Models describing the technological and human behaviour factors which influence the benefits of automation to task performance will be examined. Concepts including workload, fatigue, and trust in automation are described, as these are examined throughout this thesis. An additional consideration is then introduced: 'what happens to the human and their task performance when automation fails?' Automation is rarely perfectly reliable, either due to flaws in its design or unexpected circumstances it is faced with. The different ways automation can fail, which modulate the performance outcomes, will then be discussed. Thus, the automation section outlines various benefits and costs associated with automation use in complex task environments.

In the third section, the concept of situation awareness (SA) is introduced as it is a critical factor in maintaining safety standards and optimal task performance in environments in which automation is used. As with automation, this section will define SA, describe theoretical models,

and provide background as to its practical importance in circumstances of automation failure. Methods of measuring situation awareness are discussed as they are relevant to the methods chosen in the experiments presented in this thesis.

The fourth section will combine relevant aspects of the background literature and expand on it by discussing what is currently known about the interaction between multi-tasking ability and automation in empirical studies. These few studies are the most relevant to the aims of the current thesis and include examination of individual differences in multi-tasking-related cognitive abilities and human-automation teaming outcomes (i.e., task performance, SA, and workload). These provide a basis for the concepts and underlying assumptions investigated in this thesis.

The final section will provide a detailed overview of the aims and methodologies of the experimental chapters of this thesis.

Cognitive abilities and complex task performance

The modern workplace often requires engagement with complex tasks. Task complexity increases with the number of tasks and decisions made, time constraints, interruptions, competing priorities and uncertainty (Chérif et al., 2018; Wong & Seet, 2017). Investigating why people perform differently on a complex task is a crucial question in human factors and cognitive psychology. The study of individual differences in cognitive ability, and their effect on human task performance is a diverse area of literature. This section will describe and define multi-tasking specifically as a cognitive ability necessary to perform complex tasks and will provide a review of theoretical models of multi-tasking to explain why multi-tasking is challenging. Finally, different approaches to measuring multi-tasking will be discussed as these inform the methods chosen for this thesis.

Multi-tasking: what is it and why is it difficult?

It should be noted that 'pure multi-tasking', the concept of simultaneous attention and awareness of two sensory inputs from the same modality (i.e., auditory, or visual) is structurally limited in the brain (Duncan et al., 1997; Marois & Ivanoff, 2005). In light of this, this thesis adopted a definition of multi-tasking (also known as 'task switching' in the literature) as "the strategic direction of attention" among multiple tasks (Gutzwiller et al., 2019, p. 197). Multi-tasking has been hypothesised to be an emergent cognitive ability that is multi-dimensional (Redick et al., 2016). Indeed, multi-tasking has been associated with stable individual differences in cognitive capacities, including working memory (defined by Baddeley and Hitch, 1974 in Hambrick et al., 2010, p. 1151 as "a system responsible for both information storage and processing in the service of complex cognition", also see Redick et al., 2016), attentional control (the ability to switch attention between tasks; J. Chen & Joyner, 2009; J. Chen & Terrence, 2009), and reasoning (which can encompass verbal, numeric or figural/spatial reasoning and is operationalised by standardised tests such as IST 2000-R; Bühner et al., 2006). Studies of individual differences in cognitive

abilities have noted that some people are less prone to performance degradation during multi-tasking conditions (Rubinstein et al., 2001; Schumacher et al., 2001).

Chérif and colleagues (2018) summarised the three key constituent processes which have been theorized to comprise multi-tasking: performing multiple tasks (called 'dual-tasking'), consciously shifting between tasks (called 'task switching') and task interruption (also called 'task interference'). Each of these sub-processes are associated with cognitive limitations which result in time and performance costs. The combination of such processes and their associated limitations informs understanding of the difficulties inherent in multi-tasking. Therefore, each of these processes will be briefly reviewed here as well as common paradigms for measuring their effects on cognition, before turning to overarching models explaining the cognitive limitations underlying multi-tasking performance.

Dual-tasking: Dual-tasking is described here briefly, as the following section on cognitive response selection bottleneck will outline the costs and theoretical explanations in more detail. Dual-tasking is the primary focus of the conceptualisation of multi-tasking in the measures used in this thesis. The Dual Response Selection Task (here on called 'Dual task') paradigm (Pashler, 1984) is a well-established phenomenon that demonstrates limited cognitive capacity. The dual-task paradigm shows that when two tasks requiring immediate response occur in close succession, there is a delay in responding to the second stimulus that is not observed when the tasks are performed separately.

Task switching: The performance of a given cognitive task and the efficiency with which it is performed results from both deliberate intentions in light of goals, and the recency, frequency and availability of alternative tasks as determined by exogenous stimulus and context (Monsell, 2003). Task switching and activating a new task involves mechanisms which have been identified as the 'task-set configuration' processes. The first of these processes is to attend to new stimuli attributes. This is followed by retrieving information from procedural working memory such as the goal state (i.e., what to do) and rules (i.e., how to do it). Lastly, processes involved in engaging different physical responses and adapting to new response criteria (Monsell, 2003). These processes result in "a long term as well as a transient cost of task switching" (Monsell, 2003, p. 135). The switch cost occurs when there are greater delays in response production when switching to a new task (i.e., between trials) than for continuing to perform the same task between trials. Task switching also results in a higher error rate than remaining on the same task. The switch cost can be reduced (up to 600ms; Monsell, 2003) by allowing time to prepare for the upcoming task (i.e., by cueing the task requirements), although switching costs is not eliminated entirely (called the 'residual cost'). Other phenomena associated with task switching costs include 'cognitive tunnelling' in which humans continue to focus on the current task to the exclusion of other more important ones (Gutzwiller et al., 2019), and the 'post-completion error' phenomenon wherein switching to a new task, a person may forget to go back to complete the old task (Di Nocera, et al., 2006).

A paradigm for measuring task-switch can include the Psychological Refractory Period task (PRP; Welford, 1952). The refractory period occurs for temporally adjacent stimuli presented in the same (e.g., both visual) or different modalities (e.g., visual, and auditory). For example, when presented with a symbol for 200ms followed by a sound after 300ms, a person might take longer to respond to the sound than when only the sound is presented after a longer delay (e.g., 800ms). Thus, multi-tasking stretches limited cognitive resources under several theoretical accounts, resulting in 'cognitive overload', which costs complex task performance.

Two common explanations for the switch cost have been proposed. The cost may reflect preparatory reconfiguration, or the time to activate the relevant goal information (i.e., updating declarative memory where task demands are stored) and the time to activate the rule associated with the task (i.e., procedural memory; Rubinstein et al., 2001). This explanation focuses on consciously controlled processes such as recall (de Jong, 2000) and can be demonstrated by the reduced switching cost associated with increased preparation time (Meiran et al., 2000). A second explanation is that the processes are not under intentional control but reflect proactive interference from the previous task, called 'task set inertia' (Allport & Wylie, 2000). Thus, task switching is central to multi-tasking, as when more than one task is engaged in a period of time, there is inherent switching between them. Task switching literature helps explain how, when, and why such switches are initiated and the costs to primary task performance which result from performing multiple tasks.

Task Interruption: Task interruption (or interference) occurs when a primary task is ceased after an alert is registered for a secondary task which is then commenced. Task interruption is inherent in multi-tasking as it explains delays caused between the ending of one task and the beginning of another task (and then, the return to the original task). The sequence of events involved in task interruption is described by Trafton and Monk (2007) as the following: a person is working on a primary task (e.g., writing an email), they receive an alert for a secondary task (e.g., phone rings), they begin the secondary task. At the end of the secondary task sometime later, the primary task is resumed. During this sequence of events there are two durations of interruption which can cause delays in continuing with the primary task. These are first, the 'interruption lag' which occurs between receiving the alert for the secondary task and commencing the secondary task. Second is the 'resumption lag' which occurs between the end of the secondary task and the resumption of the primary task (Trafton & Monk, 2007). The resumption delay includes time to reorient and recall to the goals of the primary task and the next required actions. This delay can be longer if changing circumstances in the environment since the primary task was stopped require adaption or new planning. Thus, variation in the time taken to recognize and commence the secondary task and the time taken to resume the primary task may determine how much time is taken to return to a primary task (the 'interruption cost').

When performing a complex task (as opposed to a simple task) interruptions can exceed the cognitive capacity of an individual (Ratwani et al., 2006). Interruptions can have significant

costs, including increased time to complete a primary task, higher errors, as well as stress and anxiety (Adamcyzk & Bailey, 2004). These costs can come from multiple sources. Increased arousal caused by interruptions (Trafton et al., 2003) can mean relevant cues are ignored affecting primary task completion time and accuracy (Ratwani et al., 2006). Interruption costs are also modulated by the urgency of the interruption (i.e., can the person reach a logical stopping point in their primary task or must they attend to the interruption immediately). The need to interrupt one task to complete another can also lead to interference between tasks, which is particularly relevant when multi-tasking/dual tasking. This interference stems from shared demands on limited-capacity attentional resources (Chun & Potter, 2001).

A representative example of dual task inference is the Attentional Blink (AB) paradigm in which two targets are presented in a rapid serial presentation separated by distractors (Broadbent & Broadbent, 1987; Raymond et al., 1992). The AB effect illustrates that the encoding delay as accuracy (thus perception and encoding) for the first target is often high, while the second target following the first at a lag of 300ms is very poor. Longer lag delays (up to 800ms) considerably improve second target accuracy, thus providing a range for the bottleneck delay in encoding (Dell'Acqua et al., 2009; for review see Jolicœur et al., 2001), which has been found across visual and auditory modalities (Arnell & Jolicœur, 1999).

There are several theoretical explanations of task interruption costs. Prospective memory, or the intention to perform an action in the future (Brandimonte et al., 1996) has been posited as one explanation, as a prospective–task memory is created when a task is interrupted. External visual cues to aid prospective memory have been found to reduce interruption time (Dodhia & Dismukes, 2005). Long-term working memory may also be involved as interruptions have been found to disrupt encoding processes in memory (Oulasvirta & Saariluoma, 2006). Finally, the goal activation model suggests interruptions affect the activation of memory items (see Anderson & Lebiere, 1998). This model posits that memory processes which returns the most active item relevant to the current situation is central to the effect of interruption and recovery from it (Trafton et al., 2003). The model makes three predictions to explain the cognitive cost of interruptions. First, goals decay in memory which is affected by the duration of the interruption, second rehearsal (i.e., practice or reminding of goal states) can reduce the cost of the interruption, and third, cues in the environment when returning to the primary task can also reduce the cost of interruptions. These interruption costs reduce multi-tasking efficiency by the delays caused between tasks.

Theories of multi-tasking

Research on cognitive capacity limits, such as the ones described in task switching, interruption and dual-tasking literature are at the very heart of modern cognitive psychology. Early examples of research in this area include dichotic listening (Cherry, 1953), own-name phenomenon (Moray, 1959; Wood & Cowan, 1995), and filtering (Treisman, 1960), thus demonstrating a long

and continued history of investigating cognitive capacity limitations which inform modern theories of multi-tasking. Two well-established cognitive theories which directly relate to the tasks subsequently used in this thesis have sought to explain difficulties in multi-tasking: the Multiple Resource Theory (Wickens & Boles, 1983), and the Cognitive Bottleneck Theory (Pashler, 1984).

Multiple Resource Theory (MRT) assumes that performing tasks with similar resources cause tasks to interfere with each other (Wickens, 2002). Inherently, this fits with descriptions of multi-tasking as a process which relies on resources which are limited and allocatable (i.e., in the conscious control of the individual). This interference may be due to an overlap in modality and time in which two or more tasks draw on the same limited pool of cognitive resources which needs to be allocated between tasks (Wickens & Hollands, 2000). MRT is based on a four-dimension model comprising categorical and dichotomous dimensions which can explain variation in multi-tasking performance. First, there are different *stages* representing resources used for perception and cognition (e.g., working memory) which are separate from resources used in selection and response production. This means that when performing two tasks concurrently (i.e., a perceptual and a response-related task), increasing difficulty of one will not influence performance of the other. Second, perceptual modalities influence time-sharing, as performance *across* modalities (i.e., a visual and an auditory task performed concurrently) results in better time-sharing than *within* a modality (i.e., two visual tasks performed concurrently). Thirdly, the processing of focal (i.e., foveal – detail and pattern recognition such as reading) and ambient (i.e., peripheral – orientation and motion) visual channels may tap separate resources and are associated with different brain structures and information processing channels. Lastly, the processing codes dimension draws distinction between spatial and verbal/ linguistic processes, as these depend on separate resources located in different cerebral hemispheres (Wickens, 2002). Each of these dimensions describe processes or categories of resources which influence outcomes of time-sharing (i.e., multi-tasking). Thus, MRT provides an explanation for why performing certain tasks concurrently (e.g., two visual tasks in a dual task paradigm) or rapidly switching between tasks (e.g., two perceptual or two response tasks in a task-switch paradigm) result in costs to accuracy and response time compared to when completing one task at a time, thus demonstrating cognitive limitations of underlying processes. By contrast, other tasks which differ on the dimensions discussed above can be performed together without significant costs.

The 'Central Cognitive Decision and Response Selection Bottleneck' posits there are constraints on cognitive capacity and performance when doing two tasks at once (for a review, see Schumacher et al., 2001). This is based on the Multiple Resource Model (Navon & Gopher, 1979), and according to the response selection bottleneck model certain information processing stages cannot be performed simultaneously with multiple stimuli input but must be performed in sequence. Essentially this model posits three stages to information processing: perception, response selection followed by response production. When two stimuli are presented simultaneously, both are

perceived, but then one stimulus is subjected to response selection while the other remains at the perception stage. Once the response production stage has been initiated for the first stimulus, the second stimulus moves on to response selection and response production, which can occur while the response production for the first stimulus is taking place. The delay in response selection for a second stimulus may be because the two stimuli use similar stimuli-response selection processes, which causes interference and divided attention (Pashler, 1984). Converging evidence for the bottleneck from chronometric studies indicates a cognitive rather than sensory-motor delay, as the selection of the second response occurs serially after the selection of the first response. However, the second response selection can occur before producing that first response (Pashler, 1992). Other studies have confirmed that cognitive operations associated with an initial task can interfere with response selection for a second task and vice-versa (Jolicœur & Dell'Acqua, 1999). Thus, the response selection bottleneck model provides an explanation for the difficulties of performing multiple response selection tasks concurrently.

A response bottleneck is not the only explanation for the difficulties humans encounter when multi-tasking. A further bottleneck has been described in the encoding processes, much earlier than response selection, resulting in additional slowing when performing two tasks simultaneously. The encoding bottleneck described by Jolicœur and Dell'Acqua (1999), amended the previous cognitive bottleneck model by including several stages and was later expanded into the Central Capacity Sharing Model (Tombu & Jolicœur, 2003). First, the sensory encoding stage, followed by second perceptual encoding stage (i.e., recognising features) – processes combined as 'perception' in the previous model. The third stage is selective control, essentially decision making whether information needs to be remembered or ignored. If the information is relevant, it is stored in memory (called 'short term consolidation') as without this stage perceptual information degrades in 1-2 seconds and is not stored. It is these selective control and consolidation stages which cause an additional bottleneck delaying response selection for a concurrent auditory task. Validation studies of this model showed shorter response selection delays for the concurrent auditory task if the visual task required one letter to remember versus three letters (i.e., a longer delay; Jolicœur & Dell'Acqua, 1999). The last stage as per the previous model is response selection, which entails a separate bottleneck relating to single-channel mechanisms (described above).

Thus, cognitive bottleneck models describing delays in encoding and response selection processes and the MRT help explain the cognitive limitations inherent in multi-tasking. Tasks including the dual task paradigm, AB, and PRP tasks tap into different processes involved in multi-tasking including task switching, dual-tasking and rapid task switching. All these tasks stretch cognitive capacities to demonstrate the limitations of these processes which result in costs to response time and accuracy which underlie the difficulties in multi-tasking more broadly.

How can multi-tasking be measured?

There are two broad approaches to measuring multi-tasking in the empirical literature: the direct and indirect approaches. This section will briefly describe these measurement approaches with examples and outline benefits and limitations of each. The first approach is to directly investigate multi-tasking by providing situations where multiple tasks are performed concurrently (Hambrick et al., 2010). Typical performance measures including response time, accuracy, workload, and strategy, act as indicators of multi-tasking ability. There are two types of multi-tasking 'situations' that have been used. The first are short, lab-based, and typically involve task switching by presenting several tasks simulations on one screen. Examples are the SynWin system (Elsmore, 1994) which tests maths, memory search, visual monitoring, and acoustic monitoring in separate windows on the same screen, requiring participants to switch between tasks rapidly. Similar is the Multi-Attribute Task Battery-II (MATB-II: Comstock & Arnegard, 1992; Santiago-Espada et al., 2011) which comprises systems monitoring communications while also conducting tracking and resource management (e.g., fuel level). Another related task is the STEP program, based on SynWin (Elsmore, 1994), in which four quadrants of the screen present separate, independent tasks representing different cognitive demands. These may include memory, visual search, threshold comparison (i.e., does a gauge representing pressure reach critical threshold?) and event response (Cullen et al., 2013). Such lab-based tasks can also measure cognitive abilities (such as working memory, processing speed and visuospatial ability) related to multi-tasking (Bernhardt et al., 2016; Hambrick et al., 2010).

Short lab-based tasks which directly measure multi-tasking are a popular method of testing multi-tasking ability because they require little training and can be easily adapted (using more or fewer sub-component tasks: Hambrick et al., 2010). However, due to a high-level of abstraction, as while their component tasks may separately represent the sort of tasks performed in the workplace (e.g., monitoring fuel gauges or performing mental maths), when performed concurrently they do not reflect real-world multi-tasking conditions found in any job or workplace environment. The opportunities to manage the task requirements of a real workplace (i.e., prioritize or delegate tasks) are not possible in these simulators.

Another type of direct tasks used to measure multi-tasking which overcome the limitations of abstract lab-based tasks are real-world simulators. These include simulators of environments such as Air Traffic Control (ATC; Sethumadhavan, 2009), driving (Körber et al., 2015), aviation (Strybel et al., 2017) and military Command and Control (J. Chen & Terrence, 2009; Wright et al., 2018). Such real-world simulators provide greater ecological validity allowing studies which use them to draw direct conclusions about the effects of multi-tasking and cognitive capacity limitations in these environments on relevant task performance outcomes. Both lab-based and simulated direct measures of multi-tasking are operationalised in terms of the cognitive processes and stimuli involved in the particular task or simulator. For example, SynWin uses verbal and auditory

processing of verbal, symbolic and numerical stimuli, while the ATC involves response time, visuospatial and temporal processing of verbal and visual stimuli that is dynamic (compared to static in the SynWin; Redick et al., 2016). Thus, one limitation shared by a direct approach is that conclusions are potentially limited by the nature of the lab-based task or simulator which may exclude alternative conceptualisations of multi-tasking provided by a different task or simulator.

Alternatively, a second approach to studying multi-tasking is to examine the underlying processes which may constitute or contribute to multi-tasking ability. Such an indirect approach typically involves measuring cognitive correlates with multi-tasking such as working memory (e.g., measured by reading span and operation span tasks), non-verbal intelligence (e.g., Raven's progressive matrices; RSPM), attentional control (e.g., Stroop/ flanker tasks), and/or processing speed (e.g., pattern comparison; Bühner et al., 2006; Hambrick et al., 2010; Konig et al., 2005; Morgan et al., 2013; Redick et al., 2016). An indirect approach benefits from not relying heavily on the nature of the task or simulator as in the direct approach, as cognitive correlate tasks are more broadly applicable to multi-tasking and represent earlier stages in cognitive processes. Multi-tasking is thus operationalised as cognitive abilities rather than practical performance in a simulated environment which improves generalisability and validity. Cognitive processes which may contribute to multi-tasking can be combined using latent factors methodologies to extract the shared variance which represents multi-tasking (Redick et al., 2016). Latent factors benefit from not relying on any one cognitive process and its method of measurement to operationalise multi-tasking. However, it has rarely been used in experimental studies.

Both direct and indirect approaches to measure multi-tasking were used in the current thesis. First, I created a multi-tasking index score by extracting the shared variance from short cognitive tasks to capture individual differences in ability, such as the dual task, AB and PRP tasks described above. Second, an ATC task assessed multi-tasking outcomes in a simulated real-world environment. Thus, by measuring multi-tasking both as a cognitive ability (derived from the shared variance of its constituent processes) and the practical outcomes of that ability in a simulated real-world environment, this thesis can draw associations between the interaction of multi-tasking ability and automation which is designed to assist that cognitive ability with converging (i.e., combined measures), valid (i.e., theoretically driven) and relevant measures (i.e., real-world performance).

Automation augmenting human performance in complex task domains

Multi-tasking ability is particularly relevant in domains which employ automation. The primary benefit of automation is to reduce the amount of work a human does – the number of tasks or the amount of mental or physical effort required to perform those tasks. Thus, automation may reduce the cognitive requirements of multi-tasking for the human operator (Saqer & Parasuraman, 2014; Wickens, Clegg et al., 2015). The following sections will review how automation impacts performance outcomes by reducing multi-tasking requirements. First,

automation will be defined, including the taxonomies developed to describe the critical aspects of which tasks are automated and how much of an individual task is automated. The impacts of automation will then be reviewed, including how it benefits humans both in terms of complex task performance and by reducing stress, fatigue, and workload. As automation does not always work as intended or is not always perfectly reliable, the discussion will then turn to what happens when automation fails, including why and how automation may fail, and the impact on performance of different types of failure.

Automation: What is it?

Parasuraman et al., (2000) defines automation as technology that performs some or all the tasks previously accomplished by humans. This broad definition of automation encompasses two aspects: *which* tasks are automated and *how much* of each task is automated. As automation covers an extensive range of technologies and systems, it may be helpful to start by narrowing this definition by discussing what it is *not*. First, not all automation is *autonomous*. Autonomous systems incorporate technology that performs high-level cognitive tasks by entirely taking over perception, planning and decision making from the human (Roth et al., 2019). The crucial distinction between automated and autonomous systems is that the latter operates mostly independently of human oversight (see Kaber, 2018). Indeed, the goal of an autonomous system is to perform tasks for prolonged periods with limited-to-no human intervention. Autonomy can therefore be characterised as the 'evolutionary endpoint' of automation (Hancock et al., 2013). However, achieving full autonomy in any system has been an elusive goal. While exponential growth in technological development means fully autonomous systems are likely inevitable from an engineering perspective (Masakowski & Creely, 2017), many philosophical and legal concerns may limit its application (see below 'why is automation used?'). To address such concerns, automating a subset of possible system functions, often to obviate the need for human intervention, is typically the medium-term goal of system developers in industries including mining, aviation, and defence. A more helpful way of conceptualising automation is to consider it a process rather than an endpoint. Automation involves allocating functions or tasks involved in a system, deciding which tasks are performed by machines and which are performed by the human (Roth et al., 2019). This will be discussed in more detail in the review of taxonomies below.

Where is automation used?

The term *automation* refers to a broad scope of technology and systems and is therefore ubiquitous in the modern workplace. Automation technologies can be physical, such as robots, autonomous vehicles, or software systems. A computer can also be considered automation as functions (i.e., the collecting and organising of information) are performed by a machine rather than the human. At the highest end of a technology continuum, autonomous systems include artificial intelligence and machine-learning algorithms (Endsley, 2017a). At the lowest end of the technology

continuum are unaided humans. Industry employs automation for various reasons, including increasing efficiency, reducing human workload, improving safety, and reducing human error (for review see Onnasch, Ruff et al., 2014). These theorised benefits of automation have been studied across a wide range of industries that employ automation. Literature that has studied these industries has helped develop better models to understand how humans interact with automation to produce these desired – and many undesired – outcomes. Transportation automation includes railway automation in the U.K. (Balfe & Sharples 2015), automated assistants to airline pilots (Cak et al., 2020; Nguyen et al., 2019; Strybel et al., 2016), ATC (Sethumadhavan, 2009; Tran Luciani et al., 2019); military aviation (Barron & Rose, 2017; Lyons et al., 2016), and the control of Uncrewed Aerial Vehicles (UAVs; Calhoun et al., 2009; Deng et al., 2020; Freedy et al., 2007). Most recently, in road transport, driverless cars and autonomous trucks have come into mainstream use, and the impact of such automation is now being studied (Burns, 2018; Dikmen & Burns, 2016; Endsley, 2017a; Körber et al., 2015; Strand et al., 2014). Automation is also used in white-collar workplaces, such as automated financial trading algorithms that allow large-scale trading to occur in second or millisecond timeframes (Li & Burns, 2017) and in healthcare and medicine, such as assisting surgery (Manzey et al., 2011).

Why is automation used and what are the ethical considerations?

Before examining theoretical taxonomies of automation, it is necessary to address a common question regarding the future of automation in the workplace; 'why study human-automation teaming if such technology may ultimately replace humans entirely?' The possibility of automation error is a critical consideration when implementing automation. Automation may make errors or malfunction due to unforeseen technical problems caused by software bugs, hardware breakdown, deliberate sabotage (see 'cost of unreliable automation' below; Sebok & Wickens, 2017) or unexpected changes in the environment or task context. Furthermore, the dynamic nature of the environments in which automation is often deployed may mean it faces operational conditions not foreseen by its designers. Thus, automation may fail in situations it was not designed for. Automation may also be subject to *rigidity* – functioning as its designers intended but not as its operators expect, particularly in novel situations (Parasuraman & Wickens, 2008). Thus, human engagement is still required in safety-critical roles as the cost of failure in human life is high.

Ethical and legal issues that arise are additional considerations. For example, issues around liability and responsibility when decision-making automation is used in healthcare and military environments (Masakowski & Creely, 2017), such as 'who is responsible if automation makes an error?' (i.e., the designer of the system, the person in the field, or the hierarchy of decision makers who deployed it), 'will humans maintain right of supervision over autonomous systems?', and 'how does that supervision scale up to multiple autonomous systems such as swarms of UAVs?' The legal question of liability is one currently being tackled by autonomous car manufacturers and

legislatures around the world. Human drivers have a duty of care set out in the law and can be held responsible if they fail to meet that duty (i.e., if they negligently cause damage with their vehicle). When automation replaces the human driver, there is a question of whether automation has a duty of care and can be negligent in that duty. Even with semi-autonomous cars, issues arise around whether the human has the responsibility to resume manual control (Inners & Kun, 2017). For these reasons, the combination of humans and automation is essential in many settings. Hence, humans are an integral part of systems that require some automation, and the study of their interactions with such technology will continue to be relevant into the future, even when faced with the increasing technological capabilities of autonomous systems.

How is human-automation teaming conceptualised? A brief history of taxonomies.

This section provides a historical perspective on the development of taxonomies describing human-automation interactions. Developments of these taxonomies and the major criticism of each which informed the subsequent taxonomies are presented chronologically here. The earliest framework to decide 'who does what?' began in 1951 with the Fitt's List developed by Paul Fitts and colleagues for ATC and navigation systems, which already employed automation for some tasks (Fitts, 1951). The first functional allocation framework was developed from the Fitt's List. The 'Men-are-better-at/Machines-are-better-at' framework constituted a table summarising activity that are more suited to humans or to machines. The legacy of this approach is that most subsequent frameworks start by considering the strengths and weaknesses of both the human and the technology to determine which should perform what roles (Roth et al., 2019).

Levels of Automation: The roles assumed by automation can be classified by the type of task performed (e.g., ATC or medical warning systems) and the degree of control exerted by the automation (from *none* – complete human manual control, to *full automation* – no human input). The following is a brief overview of how these two concepts have developed over time, which is relevant to the discussions of the empirical findings in the literature which follows. An early classification system for whether tasks are under the control of a human or machine was the Levels of Automation (LOA) taxonomy (Sheridan & Verplank, 1978) which describes automation performing tasks of increasing complexity with greater independence. These levels have been illustrated (see Table 1.1) with general examples in a simplified model by Parasuraman and colleagues (2000). Building on Fitt's List, the LOA proposed a scale from 1 (indicating low-level automation) to 10 (high-level automation). Each level of automation depends mainly on the type of system or task (Endsley & Kaber, 1997). Sheridan and Verplank's level's taxonomy was built on by later taxonomies which described levels more generally which were more broadly applicable in a range of workplaces (Endsley & Kaber, 1997).

Table 1.1. Levels of Automation described by Sheridan and Verplank (1978) presented in a simplified model by Parasuraman et al., (2000)

High	10.	The computer decides everything, acts autonomously, ignoring the human.
	9.	Informs the human only if it, the computer, decides to
	8.	Informs the human only if asked, or
	7.	Executes automatically, then necessarily informs the human, and
	6.	Allows the human a restricted time to veto before automatic execution, or
	5.	Executes that suggestion if the human approves, or
	4.	Suggests one alternative
	3.	Narrows the selection down to a few, or
	2.	Offers a complete set of decision/ action alternatives, or
Low	1.	Offers no assistance: a human must take all decisions and actions

Criticisms of these previous automation taxonomies will briefly be reviewed here (see Kaber, 2018 for extensive review). The levels of automation taxonomies have been criticized for varying levels of inconsistencies between similar taxonomies. As noted by Beer et al., (2014) in their extensive review of human-automation teaming literature, some taxonomies (Sheridan & Verplank, 1978) focus primarily on *output* of automation stages (e.g., decision selection – what the automation *does*) while others (Endsley & Kaber, 1997) focus on *input* of automation (either human or automation providing information which the system then uses to generate options or strategies). Only describing the outcome of automated systems limits understanding of how such outcomes are reached, as it disregards how the raw data is acquired and processed. Conversely, focusing only on the system inputs limits understanding of the purpose of automation which is the practical application of such information. While both should be accounted for in a comprehensive automation framework, few of the taxonomies reviewed adequately describe the relationship between the inputs and outputs. One model which does is the model proposed by Sanchez (see 'Towards a model of Human-Automation Teaming' section below).

A further criticism of LOA taxonomies is that they assume a static allocation of functions. At each level automation does one narrowly defined thing (e.g., level 4 – suggests one alternative). Such a static approach does not allow for describing automation which performs actions across several levels or performs actions at different levels when required by situational conditions.

Stages of Automation: Subsequent taxonomies described four major task categories or *stages* (Kaber & Endsley, 2004) describing what roles automated systems could perform within a task. These roles reflect human cognitive processes, such as perception, decision making, response selection or physical action. These stages of automation are:

1. Information acquisition or monitoring
2. Option generation
3. Action selection
4. Action implementation

Stages of automation were a reaction to the limitations of the previous approach of levels that described static functions. Rouse's Adaptive Automation conceptualization (1988) involved the

dynamic allocation of controls varying allocation of a task between automated and manual over time. The dynamic nature of Adaptive Automation informed the stages taxonomy allowed for *dynamic* allocation of LOA depending on task and humans' requirements (Kaber & Endsley, 2004).

A similar Stages of Automation taxonomy which has been widely used since (Parasuraman et al., 2000) proposed the following combination of variation of stages:

1. Information filtering
2. Information analysis
3. Decision and Action selection
4. Action execution

However, rather than focusing on *outputs* (i.e., decision and action selection) from the perspective of allocating technological function, this taxonomy expanded the description of *inputs*; functions preceding decision making and action, including information gathering and processing of sensory information (Parasuraman et al., 2000). The stages of automation taxonomies were similar in many ways as they represent a 'pipeline' approach based on *human* information processing and fitting those cognitive behaviours into discrete categories (Kaber, 2018). The example of automation in cars, specifically the new generation of Tesla, nicely illustrates each stage described by Muslim and Itoh (2019). In the first stages, automation is tasked with perceiving environmental stimuli, processing, and visually presenting information for humans' use. Thus, automation can extend human sensory processing abilities, such as by using night vision cameras and rear/side parking sensors. In the second stage (combining option generation and action selection in Kaber's model) automation integrates data, diagnosis and predicts future states. In a car, automation can draw the driver's attention to potential hazards, such as pedestrian detection and traffic signal displays. The third and fourth stages of both models are identical. Automation is tasked with presenting or choosing between alternative courses of action and finally executing a decided action. Decision-making systems in cars provide warnings, such as lane departure warnings and front collision warnings. Finally, action implementation can prevent inappropriate actions by a driver, such as automatic braking systems and cruise-controlled speed.

These stages approaches have been criticised as being too broad, with unclear boundaries between stages (e.g., what differentiates a high-level stage 1 from a low-level stage 2: Pritchett et al., 2014). Another criticism is the ambiguity in the definition of roughly equivalent stages (Beer et al., 2014). An example is how information is acquired in the 'monitoring' stage of Kaber compared to the clear description of information acquisition stage in Parasuraman involving sensors used to collect and register data.

Degrees of Automation: A subsequent development was the unification of levels and stages into an integrated taxonomy that orders the complexity of automated control from lowest (automation of information acquisition fully controlled by the human) to highest (automation of action implementation without human interaction). This Degrees of Automation (DOA) taxonomy

(Wickens et al., 2010) assumes that different types of automation can be described as existing on an ordinal scale reflecting how much support automation provides. Figure 1.1 illustrates the DOA in terms of increasing across levels and stages. Essentially, an increase in the DOA (representing 'more automation') results from higher levels and later/ greater number of stages implemented (Onnasch, Ruff et al., 2014). Table 1.2 summarises this approach with examples of real-world automation mapped to the stages and levels taxonomy. The authors also distinguish between Information Automation (stage 1 and 2; IA) and Decision Automation (stage 3 and 4; DA) as these represent a critical boundary where automation starts to exceed human performance (see Figure 1.1 for illustration of boundary). Higher DOA can have benefits such as significantly reducing workload (for review, see Manzey et al., 2012) and risks the loss of awareness and skill degradation, which can have significant consequences if automation fails (Rovira et al., 2002; Rovira et al., 2007).

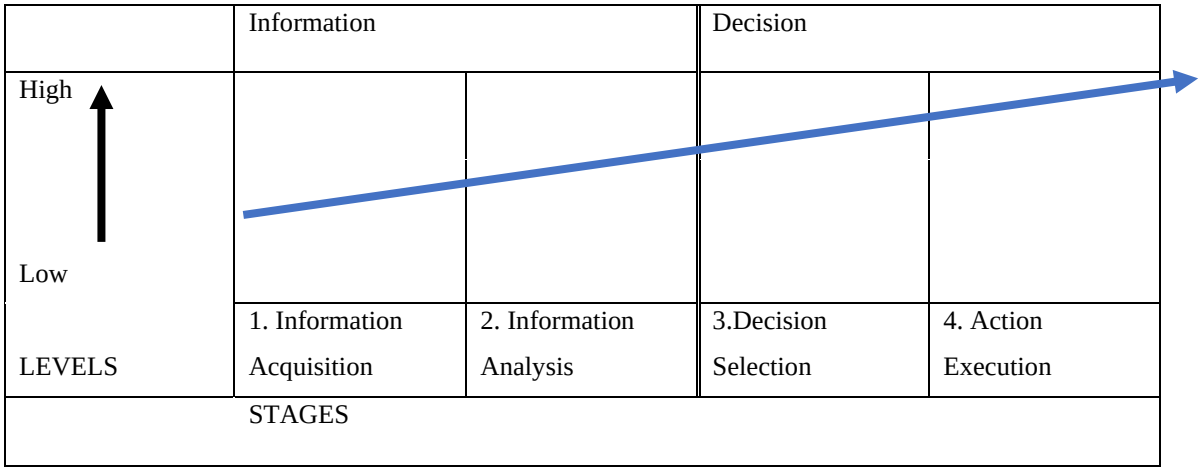


Figure 1.1. Degree of Automation: An integrated taxonomy of automation stages and levels to conceptualise 'more automation'. Blue line represents an increasing degree of automation. Adapted from Wickens et al., 2010.

Table 1.2. Examples of degrees of automation applied in real world environments. Adapted from Sebok & Wickens 2017.

	Stage 1: Information Filtering	Stage 2: Information Analysis	Stage 3: Action Selection	Stage 4: Action Execution
Example	Aircraft traffic display	Medical diagnostic assistant	Security system – intruder alert	Autonomous vehicle navigation functions
High-Level Automation	Displays only closest aircraft	Presents only most likely diagnosis	Calls police when activated	Totally automated (e.g., Google car)
Medium Level Automation	Displays all aircraft but highlights the closest	Presents all diagnoses and ranks them by order of likelihood	Calls police unless human vetoes	Lane Keeping, automatic breaking, cruise control
Low-Level Automation	Presents all near-by aircraft	Presents 10 plausible disease diagnoses	Alerts homeowner	Lane departure warning

In conclusion to the discussion of the development of taxonomies of automation, despite criticisms about the efficacy of using taxonomies to develop new systems, they have been widely accepted for describing, classifying, and analysing existing systems. Burns (2018) argues hypothesis–driven use of functional allocation models is acceptable if researchers are specific about the type, levels and/or stages of automation being used, and show how the automation is designed for the task. Such specificity will allow clear links between the automation employed in a study and the performance outcomes found, thus mitigating some unclear boundaries between LOAs depending on the task. Clarity on the nature of the automation will also allow researchers to identify the inputs and outputs present in their study which often have complex interactions with human-specific factors (trait and state based) to determine performance outcomes. This thesis adheres to this approach by clearly defining the scope and type of automation used in terms of its DOA.

Towards a model of human-automation teaming

The performance of a human-machine team can be conceptualised as the intersection of human and automation capabilities. The previous section described taxonomies focusing on automation capabilities. This section provides a similar chronological overview of models which describe how factors affecting the human in conjunction with automated processes determine system outcomes. An early comprehensive model of human behaviour in automated environments provides the basis for this perspective – overall system performance combines human accuracy, use of automation and machine accuracy (or reliability; Riley, 1996). The Riley model informed which factors specifically influence the use and misuse of automation aids by conceptualising system performance as the combination of human and machine factors. Dzindolet et al., (2001) examined automation performance outcomes in a military target detection task and from their findings, created a model grouping human-related factors into 'social, cognitive and motivational' processes. Their framework predicted automation use is based on factors including trust (i.e., social), bias towards automation and reliability (i.e., of humans' capabilities and of the automation; cognitive) and outcome values (i.e., motivational). These human-based factors contributing to automation use were mediated by external factors, including the number of tasks performed, fatigue, rewards, and interest in the task. Dzindolet's model included many aspects investigated in subsequent studies and have since become accepted aspects of human-automation systems analysis. However, this model did not clearly describe the underlying assumption of automation outcomes for performance (i.e., to be considered effective the performance of the human and automation together, called 'system performance' must be greater than either human or full automation performance in isolation).

The most comprehensive modern human-automation interaction model contains 16 variables that influence automation use relating to both human and machine capabilities and limitations (Sanchez, 2009). The Sanchez model provides a valuable starting point to describe the

factors that determine the benefits of reliable automation in human performance and overall system performance, including automation. This conceptualisation posits that there are factors (both human and automation based) that have positive or negative impacts on the use of automation. Human-based factors which have a positive impact (predict more use of automation) include perceived reliability or trust in the automation, which can be positively or negatively determined by prior knowledge of the reliability and capabilities of the system. The actual reliability of the automation also positively feeds into human's trust. Human-based factors that have a negative impact (predict less use of automation) include human's resources (mediated through workload). Strategies for the use of automation and self-confidence (their anticipated performance without automation reflecting their manual abilities) can have a positive or negative impact.

Automation or task-based factors that can positively impact automation use include cost of concurrent tasks, task difficulty, time pressure, and the number of tasks (all three of which determine task load, which positively impacts automation use through workload). The cost of the automated task reduces use, while LOA can have a positive or negative impact. All these human and automation factors that have a specific correlational impact with automation use are important to consider, as automation use itself is a critical aspect that negatively predicts overall system performance during an automation failure event (i.e., the probability of having the correct outcome and therefore overall system performance declines if automation which has a fault is relied on and heavily used).

Benefits of reliable automation: Reducing workload, stress, and fatigue

As previously noted, a primary goal of introducing automation is to augment human performance on complex tasks by reducing the multi-tasking requirements, thereby reducing the likelihood of cognitive overload. At a most basic level, reliable automation reduces task demands, resulting in improved system performance, and reduced stress, fatigue, and mental workload (Harris et al., 1995; Onnasch, Wickens et al., 2014). Each of these benefits is briefly reviewed below.

Performance: Well-designed automation benefits overall system performance (i.e., the outcome of human and machine efforts combined). When using reliable automation compared to performing tasks manually, many studies have found humans complete tasks more efficiently – including faster and more accurate performance in generic system monitoring (Cullen et al., 2014; Rovira et al., 2002; Squire et al., 2004), military Command and Control (McGarry et al., 2003; Wright et al., 2018), ATC (Strybel et al., 2016), and submarine contact tracking (S. Chen, Visser et al., 2017). Information or early-stage automation particularly has been found to result in earlier detection of errors by the human resulting in performance benefits to the system (Di Nocera et al., 2006; Rovira et al., 2002). The higher the DOA the greater performance – including accuracy and response time – improves (Wickens, Gutziller et al., 2015).

Stress: Stress impacts the efficiency of automation use and thereby system performance. Stress may be due either to the nature of the work (e.g., war fighters), environmental conditions (e.g., noise, working at night), or task demands, like performing multiple tasks at once (Sauer et al., 2012). Most types of stress lead to decrements in task performance, such as longer response times and poorer working memory, but may differ in the type of task they affect and the magnitude of the decrement. For instance, noise has been found to affect system monitoring performance more than night work (Sauer et al., 2003). Reliable automation can reduce task-related stress by reducing individual task demands or the number of tasks to bring them back within the humans' cognitive capabilities (McGarry et al., 2003). Thus, automation can have positive benefits for human well-being related to their ability to perform a task, and the requirements of that task being within the limits of their cognitive capacities.

Fatigue: Fatigue can result from prolonged manual and cognitive activity. While distinguishable from sleepiness which is related to a biological drive for sleep (i.e., circadian rhythms), fatigue lacks an agreed-on definition in human factors literature (Brandenburger et al., 2019). Fatigue has been further delineated as 'active fatigue' resulting from increased effort in interacting with a system (such as vehicle controls), and 'passive fatigue' meaning decreased task engagement resulting from assuming a supervisory role in a system (e.g., autonomous driving vehicle; Saxby et al., 2008). Both passive and active fatigue can reduce mental and physical capacity to perform a task. Reliable automation can reduce fatigue by reducing physical and mental load which prevents or reduces overload (Onnasch, Wickens et al., 2014; Parasuraman et al., 2008). Higher DOA may reduce task demands, and therefore fatigue, more than lower DOA in long-duration intensive tasks such as train driving (Brandenburger et al., 2019) and operating military UAVs (Lin et al., 2016).

Workload: Mental workload describes the cost of accomplishing task requirements (Hopkin, 1995; Stein, 1998) – specifically, the discrepancy between task demands (i.e., information needed to perform a task) and individual's cognitive resources to meet those demands (Moray, 1979). Sanchez (2009) further defines task demands as the combination of the task difficulty, time pressure and the number of tasks. Individual's cognitive resources, including perception, attention and response selection, or decision-making capabilities, may be limited (Wong & Seet, 2017) and are associated with individual differences (Redick et al., 2016).

High workload reduces task performance when task requirements exceed cognitive capabilities (Manzey et al., 2012; Onnasch, Wickens et al., 2014). Common compensatory strategies used by humans to deal with high workload can result in decrements in performance. Such strategies include intentionally lowering performance, satisficing (i.e., adopting the easiest strategy of task management) and neglecting lower-priority tasks in favour of higher-priority ones (Parasuraman & Manzey, 2010). Thus, high workload negatively impacts performance due to two interrelated factors: cognitive overload and poor compensatory strategies. By contrast, low

workload (i.e., when cognitive capabilities greatly exceed task requirements) underutilises cognitive capacities and has been linked to complacency and inattention in multiple-task environments (Parasuraman & Manzey, 2010). The consequences of high or low workload in complex task environments can be significant. For example, a 2006 review by the U.S. Army Combat Readiness Center indicates 80 to 85% of military accidents are caused by human error directly attributed to cognitive factors (Thomas & Russo, 2007).

In light of these points, of the three ways automation benefits the human discussed here, reducing workload may be the most safety critical. Higher DOAs have consistently been found to lower workload (Wickens et al., 2010; Wright et al., 2018), with one extensive meta-analysis noting as DOA increases, workload is reduced (Onnasch, Wickens et al., 2014). Lower workload, in turn, has been linked to improvements in routine task performance (Wright et al., 2018), including better accuracy (Manzey et al., 2012), faster response times (Wright et al., 2018), and greater self-rated perceptions of accuracy (Calhoun et al., 2009).

How to measure workload: Measuring workload is a contentious topic, as it can refer to different things. It can be what a person thinks they are doing, what they are actually doing, or how what they are doing relates to their capacity; each of which are measured by different methods (see Matthews et al., 2001 for review). Workload can be measured subjectively, objectively, through task performance or via physiological methods. These broad categories of workload measures will be briefly outlined below with positive and negative aspects of each also discussed.

Subjective workload often involves self-assessment, and assumes individuals have the self-awareness required to provide accurate judgements of their workload (i.e., what they think they are doing). This measurement aligns with multiple resource theory (Matthews et al., 2020; Wickens, 2002) and thus overlaps with multi-tasking. Subjective workload approaches can be unidimensional, meaning workload is conceptualised as a single outcome to measure (for review see Longo, 2018). Examples of unidimensional subjective workload measures include the Bedford Scale (Roscoe & Ellis, 1990), the Subjective Workload Dominance Technique (Vidulich et al., 1991) and the Rating Scale Mental Effort (Zijlstra, 1993). Subjective workload can also be multi-dimensional in which workload is conceptualised as originating from several possible sources. Multi-dimensional measures aggregate scores across proposed dimensions or sources of workload to create an overall index (Longo, 2018). Common examples include the NASA-Task Load Index (NASA-TLX; Hart & Staveland, 1988), Subjective Workload Assessment Technique (SWAT; Reid & Nygren, 1988), and the Workload Profile (WP; Tsang & Velazquez, 1996).

Subjective workload measures have demonstrated high sensitivity and diagnostity (Rubio et al., 2004) and may be most appropriate in many domains (e.g., Hart & Staveland, 1988, Battiste & Bortolussi, 1988). Convergent validity of multi-dimensional measures overall index scores appears to be high and highly correlated between the NASA-TLX, WP and SWAT. Concurrent validity (i.e., correlation with task performance) has also been found to be high, particularly for the

NASA-TLX (Battiste & Bortolussi, 1988; Rubio et al., 2004). The drawbacks are common with many subjective measures in general such as unverifiable introspection which may reduce accuracy or reporting reliability (Matthews et al., 2020). It is also difficult to make comparisons between different individuals on absolute scales (Longo, 2018) as it is difficult to attach unequivocal numerical values to conscious experiences (Annett, 2002). These measures are often administered post-task which influences reliability in long tasks. Finally, subjective measures have been found to be more sensitive to the task (e.g., number of tasks performed) rather than to demands of resource competition or response execution (Vidulich & Tsang 2012; Yeh & Wickens, 1988).

Objective workload often is measured by or related to task performance, although this is not the measure in this category. Task performance can be categorised into primary and secondary task assessment. Primary task performance is commonly measured by response time, accuracy, error rate, and signal detection performance (Longo, 2018). This measure has several benefits; it is a direct index of performance which can be measured over long periods thus providing a detailed picture of workload and its changes over time. It is also useful for examining workload as what an individual actually does (as opposed to what they think they do). Performance can be used to examine individual differences in resource competition. One theoretical drawback of this approach is that within the resource/demand model it is preferable to treat workload as a separate outcome that was expected to relate to primary task performance under certain conditions (Hart & Wickens, 2010). Other drawbacks are that if multiple tasks are performed by an individual concurrently, individual task performance (and therefore workload) cannot be distinguished. Often, performance is not a reliable measure of workload when used alone but can be valuable when combined with subjective ratings (Longo, 2018; Matthews et al., 2020). Further, using task performance as a measure of workload assumes that workload is only of importance to investigate if it influences performance, it does not differentiate workload as a unique outcome.

Secondary task performance can also be used as a measure of workload when such tasks add to cognitive load. Spare cognitive capacity, inversely related to the proportion of resources spent on the primary task (Gawron, 2008), can thus reflect workload by measuring secondary tasks (i.e., those tasks with no direct importance to the main task objectives). This has the benefits of quantifying spare attentional capacity and is useful for short periods of workload (Longo, 2018). Drawbacks of secondary task workload are that large changes in workload are required to detect changes, they may be intrusive and can introduce distraction by influencing behaviour on the primary task (Longo, 2018; Matthews et al., 2020).

Physiological methods are the third category of workload measurement techniques but will only be briefly discussed here as they are not part of this thesis's methodology. Such measures include blood pressure, heart rate, blinking, pupil dilation, electroencephalograms to measure brain activation signals and electromyograms to measure muscle signals (Charles & Nixon, 2019). The benefits of physiological measures are that they have high measurement sensitivity as they provide

continuous data at regular intervals. The major drawback is they are difficult to interpret and require experts and complex data analysis. It is also difficult to differentiate task-related changes from external interference (Charles & Nixon, 2019; also see Mansikka et al., 2018).

What can go wrong? Automation failure

While reliable automation has demonstrated many benefits, in reality perfectly reliable automation is extremely rare. Few systems are faultless, with only the most uncomplicated automation being perfectly reliable. Automation failure in the real world can encompass events including "software bugs, hardware failure, circumstances for which automation was not designed, and instances in which the automation performs as the designer intended but not as the user expected" (Sebok & Wickens, 2017, p. 191). The first half of this section will provide an overview of theoretical frameworks to understand the reasons for automation failure, experimental conceptualisations of failure in simulated real-world environments, followed by a brief discussion of how automation failures also impact performance.

Reasons for failure: Brittleness and rigidity

There are two broad reasons for automation failure: brittleness and rigidity. Brittleness describes situations where automation encounters circumstances for which it was not designed, resulting in incorrect or inaccurate action or decisions (Ockerman & Pritchett, 2002). These situations may encompass deliberate attacks from external forces, as in a software hack, or physical damage to the system resulting in hardware failure, both of which are instances of a system being subject to events for which it was not designed. More commonly, systems may fail to identify or misidentify events. Rigidity or inflexibility of automation occurs when automation performs as it was designed to but does not meet the operator's expectations, particularly in later-stage automation or higher DOA where systems recommend or execute an action in uncertain situations which require value judgements (Sebok & Wickens, 2017). Such differing underlying assumptions which inform the criteria used by automated systems to make decisions can result in 'wrong' outcomes or failure. Indeed, diagnostic automation is often applied in ambiguous situations (Wickens & Dixon, 2007), such as medical diagnosis. Brittleness and rigidity inform the conditions for investigating automation failures in simulated environments.

Types of failure: Misses and false alarms and their relative impact on a system

The two most common ways in which automation fails can be categorised as misses and false alarms. A *miss* is an unintentional inaction by the system, such as not alerting the human to relevant information or targets (Cullen et al., 2013). An example is the failure to detect a target in ATC (Bowden et al., 2021). Misses can lead to increased workload as humans are required to increase their vigilance to compensate for the system's failure, increasing decision-making time (Wickens, Clegg et al., 2015). In addition, due to the increased reliance placed on the automation,

misses can have a greater impact on the performance of concurrent non-automated tasks than primary automated tasks if attention is shifted to the automated task to monitor it during non-alarm periods (Onnasch, Ruff et al., 2014; Wickens et al., 2005). Systems may also fail or make an error by providing inappropriate or unnecessary alerts called *false alarms*. An everyday example of this is a smoke alarm that goes off because of a toaster or candle (Cullen et al., 2014). False alarms are common in safety critical environments, such as predictive warning systems in medical and ATC, which alert for potential future events well in advance to allow for necessary action (for review see Parasuraman & Wickens, 2008). False alarms are problematic as they may distract from useful information or performing the routine task and may erode human's use of automation that is deemed unreliable (Bliss & Dunn, 2000).

Relative comparisons of these ways automation can fail, and their negative impact have been studied. False alarms may be deemed a more serious type of failure than misses for primary task performance as they significantly impact the human and thus overall system performance to a greater extent. False alarms have significant negative impacts by increasing workload. They are more salient than misses, meaning they are recognised faster, thus eroding trust quicker and reducing reliance on automation (Wickens et al., 2005). Frequent false alarms can result in automation being ignored and potentially disregarding correct alarms for aspects of the system which are still reliable (Wickens & Dixon, 2007). In terms of the effect on system performance, false alarms have resulted in more significant decrements in accuracy than misses (Wickens, Clegg et al., 2015). In particular, false alarms can have more significant impacts on the primary (automated) task due to the effects of compliance (Wickens, Clegg et al., 2015). For these reasons, it has been argued that false alarms are more detrimental to performance than misses (Wickens et al., 2005). However, this suggestion is contentious, with one study finding misses result in poorer overall performance (Chancey et al., 2017).

The nature of performance decrements may interact with individual differences in task-relevant cognitive abilities, such as attentional control. Several related studies using low-level automation in a military Command and Control environment found those with low attentional control abilities were more adversely affected by misses than false alarms (J. Chen & Joyner, 2009; J. Chen & Terrence, 2008). In contrast, automation that yielded false alarms still assisted their performance despite the increased unnecessary actions taken (J. Chen & Terrence, 2009). Thus, the nature of the failure, the circumstances of the failure and cognitive capabilities determine its detrimental effects on performance. Furthermore, not all failure is detrimental to a significant extent as humans are resilient to a certain level of failure, applying their own skills to compensate.

Rate of failure and its impact on the system

In addition to the type or nature of the failure, the *rate* of failure or amount of error, should also be considered as a factor in determining failure outcomes. Complete failures, meaning automation goes offline and stays offline (reflecting a <90% failure rate), are easier to detect than partial failures. In simulated driving, participants have been shown to react faster and take back manual control of the system more efficiently when faced with a complete failure compared to a partial failure (Strand et al., 2014). Partial failures where a system may go offline intermittently (for discussion purposes, a failure rate of <50%, although the definition of a partial failure varies between studies) may go undetected for longer, resulting in poorer performance outcomes than complete failure (Strand et al., 2014). Between these extremes, there is a large scope of outcomes. Studies have suggested a 'tipping point' *beyond* which unreliable automation becomes detrimental to performance (Wickens & Dixon, 2007). In the seminal analysis of 22 studies, Wickens and Dixon found a positive linear relationship between automation reliability and system (joint human and automation) performance, suggesting the number of critical events detected was higher with highly reliable automation than with less reliable automation. This pattern of results only holds for automation that was more than 70% reliable; below this threshold, automation is ignored, and humans rely on their own abilities. Automation less than 70% reliable resulted in performance worse than with no automation at all (Wickens & Dixon, 2007). One study which further illustrates this threshold provided information automation and decision automation at differing levels of reliability (70%, 90% and no automation) and found performance and workload benefits from automation at both levels of reliability (Cullen et al., 2013). Thus, to examine the costs of automation failure, a minimum acceptable level of automation reliability must be applied. Automation that is less reliable than this threshold identified in the literature is unlikely to be relied on or used as intended by operators who will effectively treat the task as if it must be performed 'manually' (without automation). A summary of the effects of different rates of failure on human and system outcomes can be found in Table 1.3.

Table 1.3. Summary of effects of amount of automation failure

Failure Rate	100% (complete failure – automation no longer works)	70% (tipping point)	50% or less (unusable automation)
Outcomes	• Failure is salient – recognised quickly • Human takes back manual control, and performance improves quickly	• Humans monitor automation, • less reliant • perform better than with no automation	• Intermittent failure is less salient – impacts performance negatively for longer before the human notices • Once recognised, humans tend to ignore automation entirely and rely on their own capabilities.

Costs of unreliable automation to performance

Research has shown that costs to performance are mediated by the rate of failure (see above), as well as the nature of automation failure (i.e., miss versus false alarm). The lumberjack analogy suggests there is a trade-off between performance benefits of automation when it is reliable (e.g., reduced workload, increased accuracy, and speed for routine performance) and the potential for decreased performance when systems fail (Kaber, 2018; Onnasch, Ruff et al., 2014; Wickens et al., 2010). As the analogy goes, the higher the 'tree' (DOA), the harder the fall (performance costs). Briefly, the performance costs include degradation of manual skills, making return-to-manual performance more difficult (i.e., slower response time), and failure to detect automation errors resulting in poorer accuracy (see review by Manzey et al., 2012).

Many studies have demonstrated that higher DOA results in worse performance outcomes when systems fail (see Squire et al., 2004 for an early review; for a more current review, see Onnasch, Ruff et al., 2014). These outcomes include lower accuracy (Rovira et al., 2002) and slower response times (McGarry et al., 2003). Indeed, the costs associated with failure of decision automation assisting visual search tasks can result in performance worse than at a manual baseline (Galster et al., 2002). It has been suggested that a higher DOA makes information less accessible as humans cannot validate the decisions made by automated aids (Di Nocera et al., 2005). Operators of decision automation do not create or explore alternative courses of action or assess potential actions, leaving them at a significant disadvantage if the automation fails (Rovira et al., 2007). By contrast, information automation (low-DOA) maintains engagement and access to relevant information, thus lessening the effects of failure.

Higher workload can also be the outcome of automation failure, as in situations where the human becomes aware of a failure (either miss or false alarm) requires them to assess the automation capabilities (i.e., whether the automation is working as intended) on a moment-to-moment basis (Balfe et al., 2015). In addition, workload tends to increase further if automation that fails is at the action implementation stage or high-level (acting autonomously). Operators also need to anticipate if automation's actions must be manually corrected or intervened. Separately, a higher workload may mediate the effect of automation failure on performance. For instance, if task workload is high, automation may continue to be used and partially relied on even when it is known to be fallible, an effect well established by a meta-analysis of studies by Wickens and Dixon (2007). It has been further noted that such continued reliance under high workload circumstances is widespread in environments that require significant multi-tasking (Sanchez, 2009).

A good illustration of this comes from a series of studies by Cullen and colleagues (2013; Cullen et al., 2014) examined the effect of diagnostic (low-level, early-stage) automation on attention allocation in a multi-tasking environment (the STEP program described earlier). Participants only saw one of the four quadrants at a time but could choose to move between tasks.

Automation constituted a red border around tasks considered to be in a critical state, needing a response. Failure to attend to a critical task in the given time was considered a miss. Three conditions were included: no automation aide, 90% reliable automation, or 67% reliable automation. Automation failures could include false alarms (e.g., drawing attention to tasks that were not critical) and misses (e.g., not drawing attention to a critical task). It was found that such automation significantly affected how people allocated their visual attention, as participants' attention was allocated towards tasks based on frequency (i.e., how often a task required action) rather than criticality (i.e., if a task was in a critical state requiring action; Cullen et al., 2013). Additionally, when automation was removed, the tasks that benefitted most from automation showed the largest performance decrement (although all four tasks were affected). Workload was also assessed and found to be lower in the automation conditions than no-automation, even when at only 67% reliability.

A follow-up study by Cullen and colleagues used the same paradigm and examined experience (pre- and post-study measures of attentional allocation) and strategy (qualitative self-description of participants attention allocation strategy; coded as automation independent and automation dependent). Results showed automation benefitted participants, even at only 67% reliability, as in the previous study by Cullen et al., (2014). Experience was found only to confer benefit to efficiency with reliable automation. Analysis showed participants formed a strategy for the non-automation condition and adapted it when automation was included, suggesting automation can be seen as an addition to the base task requirements. These studies illustrate that the benefits conferred by low-level automation (even when partially reliable) can influence attentional resources, performance, and workload in a multi-tasking environment.

In summary, automation is commonly used in complex task environments such as aviation to mitigate the cognitive limitations inherent in multi-tasking, which is a central requirement of such environments. The DOA framework used to define the combination of what tasks are automated and by how much it is frequently used in automation literature as the nature of automation is a moderating factor in task performance, workload, and human (i.e., stress and fatigue) outcomes. Higher DOA is typically associated with improved benefits to performance, and reduced workload. However, automation may also not be perfectly reliable, and may fail due to issues of brittleness and rigidity. The way automation fails, either missing relevant alerts or providing unnecessary alerts (i.e., false alarms), may modulate the effect of failure on performance and workload. How often a system fails (i.e., the rate of failure) also modulates performance outcomes.

Situation awareness

Situation awareness (SA) "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1988, p. 97) which can then mediate performance outcomes. The following section will overview the psychological construct of SA, starting with definitions. Next, it will describe relevance to the domains hereto discussed, followed by a brief overview of competing conceptualisations of SA. Next, this discussion will turn to operationalising SA – what is good SA, and how can it be measured? This section will include an overview of the techniques and assessments of SA and their relative advantages and disadvantages as these inform the methods chosen in this thesis. Finally, in the last part of this section, the benefits of good SA and the performance costs when it is lost will be examined in relation to automation, with a particular focus on ATC literature.

Definitions and relevance

Various approaches to defining and conceptualising SA have been proposed, including an ecological approach (Smith & Hancock, 1995) and a perceptual cycle model (Bedny & Meister, 1999). However, Endsley's information processing-based model (1995; 2000) underpinned by cognitive theory has driven subsequent research since the 1990s (Salmon et al., 2009). An early definition of SA by Endsley, which has been subsequently expanded to the current tri-partite model commonly used in aviation and military contexts, describes SA in terms of perception, comprehension, and projection (Endsley, 1988). These factors act as a hierarchy, from a *low*-level SA (perception) to a *high*-level SA (comprehension). Each component depends on the adequate acquisition of the previous. A detailed breakdown of these levels of SA follows (Endsley, 2000):

1. Perception: basic unit of information encoded in the visual system
2. Comprehension: to create meaning, information needs to be combined, interpreted, stored, and retained for future use. Information is prioritised by subjective interpretation (based on perceptual awareness) and objective importance (based on the situation).
3. Projection: the highest level of understanding. This requires a human to anticipate future events based on their comprehension of current events.

Defining SA, and whether the term pertains to the process of attaining awareness or the state of being aware (i.e., the end product), has been a controversial issue in the literature (Rafferty et al., 2008). Some have argued that SA refers to the emergent property of a series of related cognitive processes (Sarter & Woods, 1995). The ‘process’ view suggests SA develops by interacting with a system or from cognitive processes involved in that interaction (Falkland & Wiggins, 2019). The cognitive processes involved may include perception, working memory, attention, and executive control (O'Brien & O'Hare, 2007). Put differently, this view suggests SA is the processes used to acquire understanding of a situation.

Others have argued that the set of processes that make up SA is distinct from the mental processes used to gain that knowledge and occurs before, and is separate from, decision-making and action implementation (Endsley, 1995). This view suggests SA is a 'product' that emerges from comprehension and perceptual awareness. Under this view, SA is the conscious information held in memory that can be transmitted between individuals (Endsley, 2015), incorporated into mental models (Endsley, 2019a), or used in decision making (Bakdash et al., 2020; Deng et al., 2020). Put differently, this view suggests SA is the content or understanding of what is known about a situation. The 'process' perspective is critical of the 'product' view as it fails to explain how people come to comprehend, citing the possibility of people performing a task well by 'doing the right things' without a critical understanding of their actions (Durso & Sethumadhavan, 2008). Verbatim repetition of information could produce comprehension (i.e., knowing a thing, or 'implicit knowledge') without real understanding (i.e., knowing why a thing is true or correct, or 'explicit knowledge'), which is required under the product conceptualisation of SA. These perspectives allow for a range of understandings and measurements of SA, the combination of which can produce a very theoretically robust measure of SA through both implicit and explicit assessments. The majority of SA literature, including the literature here reviewed, deals with SA as a state of knowledge (i.e., the end product: Endsley, 1995).

Beyond Endsley's Cartesian approach that locates SA in the mind of the human operator, several approaches have developed alternative conceptualisations. *Situated SA* is based on Pickering & Garrods (2004) Interactive-Alignment Model. This model states SA emerges automatically from unconscious priming that occurs when individuals interact with each other or a system which generates a shared understanding of a dynamic situation. Thus, in situated SA understanding is distributed across agents (individual or technology or combination) as a "joint cognitive system". Rather than maintaining effortful internal representations in working memory as in the Cartesian model, the individual avoids excessive working memory requirements by off-loading information to the environment (e.g., locating information externally to a computer (Chiappe et al., 2012; 2015). Situated SA holds that individuals sample limited amounts of information from the environment only when immediately required and only what is directly relevant to the task at hand. In this view, SA constitutes knowledge of goals and required actions, and the location of relevant information in the environment to complete those actions (Chiappe et al, 2012). In sum, situated SA is compatible with Endsley's product view of SA, as both suggest SA is a shared mental representation of the environment (Endsley, 1995). However, these perspectives differ in that situated SA holds individuals possess only a partial representation derived from constant interactions with the environment (Chiappe et al., 2012).

The *distributed approach* (DSA) to conceptualising SA developed by Stanton et al (2004; 2006; 2009) holds that the working memory requirements of creating shared mental models as in the Cartesian/ Product and situated SA models are too costly or not possible. DSA examines the socio-

technical system as a whole, describing transactions of information between agents (human and technological) about the environment (Stanton, 2016). Under this view it is not relevant which agent hold the information, only that information is activated and transmitted between agents when needed. DSA holds there is no 'one SA model' but rather a multi-faceted awareness of the environment held by different agents who share information when needed and relevant to their individual and collective goals. This dynamic and emergent SA changes moment to moment with changes in the environment, task goals and agent's interactions (Stanton, 2016). It is critical to note that DSA is not shared SA. In fact, DSA explicitly states SA cannot be shared, because of "the inevitable variability in goals, roles, experiences, training, knowledge, skills and attitudes across the team" (Stanton et al., 2009, p. 51). Shared SA implies common goals and understandings, whereas DSA implies different but compatible goals and individual understandings across a socio-technical system. In sum, DSA refers to information activated for a task which exists in different parts of a socio-technical system (Stanton et al., 2006).

SA is of practical relevance to safety-critical domains, including the military (Matthews et al., 2001) and aviation (Nguyen et al., 2019) and is often discussed in the context of specific occupations or systems. Therefore, SA is best operationally defined in terms of specific roles – so-called 'domain dependent' (O'Brien & O'Hare, 2007). An example is air traffic controllers who monitor aircraft flight paths, constantly updating their knowledge, and extrapolating current flight information to future states to maintain safe operations of an ever-changing number of aircraft (Endsley, 1995). SA has a long history in the aviation industry and is an integral part of pilot training (Nguyen et al., 2019). Furthermore, SA has been identified in reviews of aviation safety incidents as *the* contributing factor (Woodhouse & Woodhouse, 1995), as human error accounted for 70% of accidents worldwide (Helmreich & Foushee, 1993) and 80% of general aviation incidents (Lau, 2007). Pilot awareness was determined as contributing to 50% of fatal incidents between 1999 and 2008, which occurred during descent and landing (Boeing, 2009).

The theoretical relevance of SA in complex task domains, and the likely explanation for its critical importance to aviation, is that it has been identified as 'the major factor driving the quality of decision process' (Endsley & Jones, 1997). Others have described it as the basis for 'effective choice' (Wong & Seet, 2017). SA, therefore, relates to the cognitive faculties such as decision making, perception and multi-tasking, which are central to effective task performance in automated domains (Redick et al., 2016). SA is particularly relevant when automation is imperfect as it may mediate between task performance and type of automation (Sethumadhavan, 2009), such that poorer SA was associated with poorer task performance when automation is unreliable (Wickens et al., 2010).

How to measure SA?

The difficulty in defining and measuring SA has been well documented. Many studies which compared common SA measurement questionnaires found they were not correlated (Cak et al., 2020), indicating fundamental differences in the underlying construct being assessed. Construct validity depends on appropriate measurements methods which are underpinned by the theoretical perspective of SA that is used. For example, the *product* view of SA is best measured with memory-based methods; these can include the Situation Awareness Global Assessment Technique (SAGAT: Endsley, 1995). By comparison, *process* SA and situated SA can be measured using response time latency and accuracy to infer implicit cognitive processes by measures such as the Situation Present Assessment Method (SPAM; Durso & Sethumandhavan, 2008). Distributed SA can be measured using network analysis (Sorensen & Stanton, 2016), but as this method does not fit into the classifications described below it is beyond the scope of this thesis.

Despite the philosophical issues regarding the validity of SA measures resulting from the conflicting definitions (e.g., 'is there an 'ideal' awareness of a situation?', 'can different people form the same awareness of a situation?', and 'is there an objective reality to compare awareness to?'), construct and face validity can be meaningfully assessed (Salmon et al., 2009). There have been a recent meta-analysis (Endsley 2019a) and reviews on SA measurement (Endsley 2019b; Kaber et al., 2006; Nguyen et al., 2019) along with direct empirical comparisons of SA measures (Salmon et al., 2009). Broadly, there are two categories of measurement tools: objective and subjective assessment. These can be further classified into six types of SA technique (Nguyen et al., 2019). What follows is a brief description of each technique. The relative strengths and weaknesses, as well as examples, are summarised in Table 1.4.

Subjective measures: Assessment of SA may be obtained by either direct questioning of the human or by independent observer ratings (Endsley, 2019a). Subjective measures may use a Likert scale, or similar, to probe a person's assessment of their own SA (Endsley, 2019b). The Situation Awareness Rating Technique (SART) is the most common (Taylor & Selcon, 1990) of these techniques and asks questions about understanding, supply, and demand of attentional resources. SART is administered post-task and focuses on generic, global aspects of the scenario to derive a subjective measure of participant SA, such as situation instability, complexity, changeability and participant alertness, concentration, knowledge, and experience (Rafferty et al., 2008). SART has not been found to correlate with objective measures such as the Situation Awareness Global Assessment Technique (SAGAT: Endsley, 2019a).

Observer ratings are commonly used in real-world task environments where an unintrusive in-situ assessment of SA is required (Salmon et al., 2009). An example is the Situation Awareness Behavioural Rating Scale (SABARS) used in the military in field-training exercises (Matthews & Beal, 2002). This method commonly requires an expert to determine observable task-relevant behaviours and assess the participant performing the task on a five-point rating scale.

Objective measures: Situation Awareness Global Assessment Technique (SAGAT) is a commonly used objective measure which has been validated in recent meta-analysis (Endsley, 2019a). This measure involves verbal or visual queries about task requirements determined by a query development technique called 'goal-directed task analysis' or by domain experts (Endsley & Jones, 1997). SAGAT is typically used for simulated environments such as ATC (Jipp & Ackerman, 2016; O'Brien & O'Hare, 2007), military Command and Control simulators (Wright et al., 2018) and simulated submarine task (S. Chen et al., 2018). In SAGAT, the task is paused and displays go blank at intervals throughout the task while questions are presented, with participant answers compared to the current state of the environment (or 'ground truth') to determine SA (Endsley, 2000; Kaber et al., 2006). Queries assess SA across the three levels of the Endsley (1995) framework – perception, comprehension, and projection – allowing for diagnostic assessment of SA at different levels.

Situation Present Assessment Method (SPAM) measures SA by presenting queries at intervals during a task (i.e., online probes) and is commonly used in ATC task studies (Edwards et al., 2017; Keeler et al., 2015; Strybel et al., 2016). The assessment method provides accuracy and response time as indicators of SA (Durso et al., 1998). Unlike SAGAT, SPAM questions are not typically presented during a pause of the primary task, which can increase workload and require multi-tasking (Endsley, 2019b) or bias response time during low-workload periods. SPAM questions may differ from SAGAT questions, for example, in ATC scenarios, by questioning relative information (e.g., which aircraft is higher?") rather than questioning absolute information (e.g., "what is the altitude of X aircraft?": Strybel et al., 2016). In addition, queries may be in terms of 'past, present and future' states (Endsley, 2019b). As displays remain visible during the presentation of probes the SPAM measure is consistent with situated SA in that humans retrieve information from the external environment rather than represent all information in internal mental models (as is assumed when the display is not visible during probe presentation such as in SAGAT: Durso et al., 2008).

Performance measures: It has been suggested that SA is not beneficial by itself (Endsley, 1995; 2000), and having good SA only benefits when it results in good task performance (as in instances of automation failure). SA has variously been described as the precursor to, fundamental construct in or basis for performance (Bakdash et al., 2020). However, a recent review of the effect sizes associated with 38 SA studies has found that SA was associated with a small effect on performance (Bakdash et al., 2020). Performance has also been criticised as confounding experience with SA, as an expert may have poor SA but rate highly on task performance because of their learned skills, while a novice may have good SA and still perform poorly (Nguyen et al., 2019). Thus, task performance is potentially not a good indicator of SA.

Process Indices: Measures such as eye-tracking can record the processes that underlie SA or that are used to develop SA (such as attention allocation). Eye movements combined with verbal

protocols ('thinking out loud') whereby the participant narrates their behaviour to create a transcript as they perform the task. Combining these objective and subjective measures provides insight into cognition and behaviour used to reflect SA (Nguyen et al., 2019; Salmon et al., 2009).

Table 1.4. Assessment of situation awareness: Advantages and disadvantages. Adapted from Nguyen et al., 2019.

Category: Type	**Example**	**Advantages**	**Disadvantages**
Subjective: post-trial ratings	SART	• Easy to obtain • Quick to complete • Low cost • Non-intrusive to task performance (Salmon et al., 2009) • Do not require domain experts	• Humans may be unaware of their level of awareness (e.g., what they don't know) • May not be a sensitive measure of SA – more related to assessment of performance
Objective: Freeze probe	SAGAT	• Direct and objective • Sensitive to design manipulations • Predictive of performance (Endsley, 2019b)	• Intrusive, requiring a simulated task (cannot be used in-field) • Questions of validity may conflate SA with memory (Salmon et al., 2009)
Objective: Real-time probe	SPAM	• No interruption of the task – less intrusive than pause techniques • It can be used in-field with real-world tasks	• Intrusive • Questions may cue participants to relevant information in the task, biasing results • May require domain experts to create probes • Difficult to use in team tasks • Lower sensitivity compared to SAGAT (Endsley, 2019b) • May conflate SA with memory (Salmon et al., 2009)
Process Indices	Eye tracker	• Non-intrusive • Assesses perceptual SA well – that subject fixes on	• Indirect measure of SA. • Expensive equipment and difficult to implement in field environments • Perception does not equate to comprehension or projection aspects of SA (limited scope)
Performance measures	Operation score	• Easy to collect (part of existing task's accuracy and RT)	• May confound experience with SA • Low correlations were found between SA and task performance

Benefits of good situation awareness in automated task environments: Air Traffic Control studies

The evidence reviewed so far suggests that the assessment of SA is context dependent. Therefore, measuring SA requires reference to the specific domain it is being measured in. As this thesis uses the ATC task as its automation task, the following discusses the benefits of SA and its relationship to automation and task performance in studies using ATC tasks. ATC is a safety-critical environment where high-quality performance is paramount. It often involves high workload, stress and fatigue and a controller interacting with automation, making it an ideal candidate for assessing SA (Edwards et al., 2017). By way of a brief description, the ATC task provides flexibility in deploying automation, measuring SA and workload in a dynamic, multi-tasking-environment. ATC maps on to Chérif's definition of multi-tasking as it involves monitoring multiple aircraft travelling through the sector at once to perform three tasks: accept aircraft entering, hand-off aircraft exiting, and detect conflict between aircraft on intersecting flight paths at the same altitude. These tasks are performed concurrently, requiring rapidly switching between these, and thus interruption – particularly of conflict detection which occurs over several minutes. ATC was used in previous studies examining working memory, which is a constituent part of multi-tasking (Redick et al., 2016). Also, it is accessible to novices and representative of real-world tasks (Fothergill et al., 2009).

Good SA in the ATC task includes knowing the location, movement, and direction of multiple targets in space to maintain safe operations (Falkland & Wiggins, 2019). An early investigation of SA in an ATC environment with air traffic controllers compared four measures: subjective (SART), objective freeze probe (SAGAT), objective real-time probe (SPAM) and performance. Overall results indicated SA predicted performance of air traffic controllers above workload measures, and all SA measures predicted the performance evaluations of subject matter experts. However, only the objective measures predicted actual task performance (Durso et al., 1998). These findings were among the first to illustrate the unique contribution of SA and its tenuous relationship to workload.

The nature of automation (i.e., DOA) applied in the ATC environment has been found to modulate outcomes for SA, indicating SA may be augmented or reduced by automation. Studies have found low-DOA improved SA and task performance (Kaber et al., 2006), even with imperfect or unreliable automation (Sethumadhavan, 2009). A series of studies with student novice participants and air traffic controllers (Edwards et al., 2017) measured SA by response time to SPAM in the ATC task. It was found SA predicted conflict-detection performance and was maximised under low or no-automation conditions, which has been theorised to enable humans to update their picture of the environment to maintain good SA (Edwards et al., 2017). Performance was poorer for conflict-detection under the most highly automated conditions (where only conflict detection was automated and other tasks remained manual), which coincided with low workload

and poorer SA, possibly reflecting underload (i.e., cognitive resources exceeding task demands such that the human is not sufficiently engaged). Comparable results were found with air traffic controllers using NextGen ATC automation to determine which tasks should be automated and which environmental circumstances maintain optimal SA (Strybel et al., 2016). High-level automation of several tasks (i.e., aircraft spacing, the allocation of airspace to provide separation between aircraft) reduced workload and resulted in poorer SA under challenging environmental circumstances (e.g., higher than average traffic, poor weather conditions). Manual control by experts was ultimately recommended. These studies demonstrate the importance of maintaining good SA to maximise task performance in the ATC environment. SA can be assisted by low-DOA but degrades with higher DOA. Furthermore, SA is related to task performance under low-DOA but may be less so under higher DOA.

Costs of loss of situation awareness:

While good SA promotes performance benefits in complex environments, loss of SA can have equally negative consequences to performance and safety (Li & Burns, 2017; Manzey et al., 2012). Loss of SA can be classified under two types of problems: failure to detect a potential problem and failure to understand a problem (Kaber & Endsley, 2004). Failure to detect a problem, such as imperfect automation 'missing' a target, and a lack of understanding of the problem due to reduced system engagement can result in delayed response time to compensate for such automation failure (Squire & Parasuraman, 2010). When automation fails to perform as expected, and the human notices, they must first recover a sufficient state of awareness about the system before taking back manual control. Regaining control may require re-engaging cognitive task demands (e.g., what am I doing?), re-acquiring relevant information (e.g., what is the current speed and trajectory of a vehicle?) and manual skills (e.g., how do I do this?) and then performing those actions required to avert a safety incident. The resumption of SA after a loss has been described in the four-stage model by John and Smallman (2008), including real-time change detection, pre-interruption preparation, post-interrupt reorientation, and post-hoc change detection. This model posits that before an interruption, a person is monitoring for changes with the intention of maintaining SA. Pre-interruption preparation is the time a person has before a forthcoming interruption in which to prepare, with more serious SA loss results from shorter preparation time. After the interruption, people must reorient themselves to the goals and task state, involving recall or external cues. Reorienting is influenced by the amount of change in the situation and the duration of the interruption. Lastly, people must detect and comprehend the changes that occurred during the interruption.

Human factors literature often connects the loss of SA and passive monitoring of automation to out-of-the-loop (OOTL) situations where the human in control of an automated system is no longer engaged with the task (Endsley & Kiris, 1995; Kaber & Endsley, 2004). The

real-world costs of OOTL in aviation and safety incidents have been discussed earlier in this section. Three components of OOTL have been described by Sebok and Wickens (2017) as: complacency (i.e., reduced vigilance and increased trust in automation), passivity (i.e., reduced system engagement and fail to notice changes in the environment), and loss of skills required for appropriate intervention (i.e., return to manual operation).

In summary, SA is a critical factor in maintaining safety standards in environments in which automation is used, particularly aviation. SA is critical precisely because automation is fallible, meaning humans must maintain the awareness and skills required to regain manual control in the event of automation failure. It must be noted that SA is a contentious area of research, with competing definitions and methods of investigating and measuring concepts. However, it has good face validity and practical relevance. SA outcomes are modulated by DOA, such that low-DOA may augment SA, while higher DOAs reduce SA. Task performance outcomes when automation fails are modulated by SA, as maintaining good SA may reduce costs in such circumstances.

Cognitive ability and human-automation teaming

Having established the theoretical and practical importance of multi-tasking earlier and having examined literature on performance of complex task environments aided by automation, it is necessary to turn to the few studies that have examined the impact of multi-tasking ability on automation outcomes in simulated task environments. It is in the interplay between cognitive abilities and automation that we can understand how, and more importantly why performance, SA and workload outcomes occur vary. It should be noted that very few experimental studies have operationally defined multi-tasking itself, but rather have examined individual differences in related abilities, including information processing (Jipp & Ackerman, 2016).

A study that examined the association between individual differences and automation was conducted by Jipp and Ackerman (2016) using a simulated ATC task. This study examined individual differences in information processing and working memory. Automation was provided at a low-level (i.e., information analysis), intermediate-level, and high-level (i.e., decision selection). The findings support previous literature in which performance in general increased as the level of automation increased (Wickens, Clegg et al., 2015). In contrast, SA showed the opposite pattern, with the best SA supported by intermediate-level automation, poorer with low-level, and worst with high-level automation. Individual differences in the capacity to process information and working memory reflected the expected relationship to performance, with higher capacity related to better performance outcomes and SA. Similar to the findings of Wright et al., (2018), the interaction of cognitive capacity and automation showed a contrasting pattern of results. Those with the poorest information processing ability performed best with low-level automation and worst with high-level automation, whereas those with the highest capacity performed best with high-level automation and worse with low-level automation. Working memory showed the same pattern of

performance related to the level of automation and capacity. Unlike in Wright's study, where participants with the lowest capacity performed the best with automation that provided the *most* assistance, this study assessing individual differences in another cognitive ability showed the opposite preference for automation; individuals with poorer capacity performed best with automation that provided the *least* assistance. The authors speculated this was due to the limited complexity of the mental models created by those with poorer information processing abilities. The complexity of the mental model required to perform the task increased as automation assistance increased to accommodate the changes in the task requirements. Thus, those with limited mental models coped best with simple automation that did not require them to integrate new functions, which was required under higher levels of automation. By contrast, more capable participants who can create complex models of the task could adapt their understanding to include more complex automation functions in higher automation levels and thus perform better. Therefore, the interactions between system and human factors may best explain patterns in performance in simulated workplace environments.

While DOA may be one factor that can modulate SA outcomes, individual differences in human cognition have also been shown to affect SA in the ATC environment (Jipp & Ackerman, 2016). For example, working memory and information processing ability (measured via intelligence test containing spatial, verbal, and numerical components) were found to interact with DOA to predict SA (Jipp & Ackerman, 2016). Precisely, perceptual-level SA was predicted by information processing abilities. Concerning automation influencing SA, the study concluded those with superior cognitive ability who used higher levels of both IA and DA automation had better performance and SA. The authors theorised that this interaction was due to those with superior cognitive abilities being able to form more complex mental models that incorporated automation capabilities, and thus such people maximise the benefits of high-level automation (Jipp & Ackerman, 2016). These findings demonstrate the complex interactions between automation, SA, performance, and human cognitive abilities.

This study indicates, firstly, it is possible to profile individual differences in cognitive capacities related to multi-tasking and demonstrate that cognitive capacity varies the impact of automation on complex task performance (Jipp & Ackerman, 2016). Secondly, interactions between cognitive capacity and automation predict performance and SA outcomes (Jipp & Ackerman, 2016). Lastly, as discussed in the automation sections above performance also varies as a function of the level of automation, which can be modulated by the reliability of the automation (Cullen et al., 2014). These three findings set the foundation for the experiments in the current thesis.

Thesis Overview

While DOA has been shown to mediate the effects of automation on performance, an equally important factor affecting human-automation teaming outcomes is individual differences in cognitive capabilities and cognition. A limited number of prior studies have examined cognitive abilities, including attentional allocation (Cullen et al., 2014), working memory and information processing (Jipp & Ackerman, 2016) concerning task performance, SA and workload outcomes when using automation. However, these studies do not examine how individual differences in these abilities might interact and modulate the impact of automation on performance, SA, and workload. Such an outcome would be of considerable practical importance as it would imply that individuals could benefit differentially from the same automation. For example, low-DOA, which to-date has shown to be of lesser benefit to the operator than high-DOA (Onnasch, Wickens et al., 2014), may be of considerably greater value for operators with poorer cognitive abilities than currently thought. Low-DOA may be easier to implement in many workplaces than high-DOA, less costly to develop and may have lower risks associated with automation failure. Whereas earlier literature has looked at how individual differences modulated the impact of *different* DOAs *within the same individual*, this thesis will investigate whether individual differences modulate the impact of the *same* DOA *across different individuals*. Reliably demonstrating such differences is the first step to understanding why such effects may occur which may help the design of more efficient human-automation teaming systems in the future.

Thesis aims

The first question this thesis asks is 'does profiling individual differences in a task-relevant cognitive ability predict performance outcomes in a complex task?' The relevant cognitive ability chosen was multi-tasking which was indexed by combining performance across a range of short multi-tasking-related tasks. As illustrated earlier, multi-tasking is an important cognitive ability that impacts task performance in complex environments of the modern workplace. This cognitive ability differs between individuals and comprises decision making and directed attention among other related executive processes in a person's conscious control.

The second question this thesis asks is 'does using automation designed to augment multi-tasking result in better outcomes than without automation?' Automation is primarily designed to reduce task demands (i.e., the number and/or actions required for a task) and thus is related to multi-tasking.

The third and most theoretically interesting question this thesis asks is, 'do the outcomes of automation within a DOA vary with user cognitive abilities (in multi-tasking)?' If such an interaction effect was found, this would provide the first evidence of its type that automation outcomes vary with cognitive ability, such that poorer multi-taskers may receive greater benefits

from low-DOA than better multi-taskers, and that the benefit they receive is proportionally greater given the fewer cognitive resources they bring to the task. Likewise, a similar outcome may occur when automation fails. Those who received the greatest benefit from reliable automation may also suffer the greatest cost when it fails. Such an interaction was only expected if effects of low-DOA and multi-tasking were found. This has not been empirically examined in previous literature but may have significant implications for our theoretical and practical understanding of human-automation outcomes, including whether automation benefits some people more than others, whom it benefits, and what those benefits are in terms of automation and non-automation task performance, SA, and workload.

Thesis methodology

To operationalise multi-tasking, the shared variance from short multi-tasking-relevant cognitive tasks (i.e., the PRP task, Dual task, and AB task) was extracted using latent factors methodology. This approach has the advantage of operationalizing multi-tasking with respect to its constituent parts, including task switching, dual-tasking, and rapid switching, rather than relying on a single-task measure. The same cognitive tasks were used across all Chapters.

To investigate the benefits of automation and the costs of unreliable automation, the present thesis used the well-validated ATC task introduced by Fothergill et al. (2009). Participants monitored multiple aircraft at varying altitudes, trajectories, and speeds, and had to: a) accept aircraft approaching their controlled airspace, b) hand-off aircraft departing their controlled airspace, (i.e., each within a 20-second timeframe before the aircraft passed over the sector boundary) and c) detect conflicts that occurred when two aircraft violated minimum separation rules (i.e., 10,000 ft vertical and five nautical miles horizontal separation). There were 45 instances of acceptances and hand-offs each, as well as 10 conflicts in each 30-minute scenario of the ATC. Participants completed one scenario without automation and one with automation. The automation used was low-DOA which increased the saliency of events that required participant action by providing visual cues to aircraft requiring operator actions. In the case of acceptances and hand-offs, aircraft that required action changed colour and flashed (Chapters 2 and 3). For conflict detection, automation highlighted aircraft red that could potentially breach separation standards (i.e., real conflicts as well as six scripted near-miss events – Chapters 3 and 4).

One of the novel aspects of this thesis is the investigation of a single DOA to examine whether outcomes vary within rather than across DOAs as has traditionally been studied (Onnasch, Wickens et al., 2014). Low-DOA has been found to modulate performance benefits compared to higher DOAs when examined on a continuum. However, no studies have examined whether benefits may vary *within* a DOA, resulting in some people receiving significantly greater benefits than others depending on the cognitive resources they bring to the task. This is one of the aims of the current thesis.

Situation awareness was measured in all studies using a modified Situation Present Assessment Method (SPAM; Durso & Dattel, 2004). This method is well suited to examining SA in the ATC as it delivers SA queries in real time while the display remains visible (Chiappe et al., 2012; Loft et al., 2016). SPAM accuracy and response time has been correlated with performance in the ATC task (Bacon & Strybel, 2013; Durso et al., 2004) and time taken to accept queries has been positively correlated with subjective (Loft, Bowden et al., 2015; Vu et al., 2012) and objective workload (Loft, Sadler et al., 2015). In total, eighteen SPAM prompts were presented in each 30-minute ATC scenario with questions assessing awareness across past, present, and future information relevant to acceptances, hand-offs, and conflicts. An adjusted RT was calculated for SA by dividing mean RTs by mean accuracy, providing a single composite measure which combines accuracy and RT to account for speed-accuracy trade-offs (Liesfeld & Jancyzk, 2019; Visser et al., 2015).

Subjective and objective methods of workload were used to allow comparisons and provide converging evidence of the impact of automation. Subjective measures included a single-item measure (Chapters 2, 3, and 4) and the NASA-TLX (Hart & Staveland, 1988; Chapter 3 and 4) completed after each ATC scenario. Objective workload was measured by response latency to the SA prompts (Strybel et al., 2016), as responses to a secondary task reflect spare capacity (i.e., SA queries) is well-founded (Matthews et al., 2020).

The overall experimental design allows for comparisons between and within-subjects, testing for the main effects of individual differences in cognitive ability and automation, and their interactions. Such a design allows this thesis to draw conclusions about why such effects occur, which may assist the future design of automated systems in augmenting their human-automation teams.

Summary of Chapters

The first experiment (Chapter 2) asks the three questions outlined earlier to establish the automation task performance outcomes with reliable low-DOA applied to acceptances and hand-offs only. SA and workload were examined as described above. Conflict detection, which was not subject to automation in Chapter 2, allowed for an additional outcome to be examined – the benefit or cost to a non-automated task performance. This first experimental chapter establishes the methodological framework used in the subsequent studies.

The second experiment (Chapter 3) applied the methodology from the first experiment and asked the same foundation questions allowing for the partial replication and extension of the findings of the first experiment, in addition to the low-DOA applied to acceptances, and hand-offs low-DOA was applied to conflict detection. Conflict detection was a more ambiguous and challenging task in the ATC environment thus, extending the approach to this task improves the generalisability and applicability of the results of the second experiment. Improvements were

made to the SA prompts to increase their validity. Additional measures of subjective workload were added to allow convergent comparison of results and potential for replication of results from the workload measures used in the preceding study. In combination with the preceding experiment, the second experiment provides a good basis for understanding the benefits of low-DOA and multi-tasking ability on complex task performance, SA.

The third experiment (Chapter 4) applied the same methodological framework and incorporated new questions; 'what happens to performance, SA, and workload when low-DOA makes an error, and how does this interact with multi-tasking ability?' Specifically, 'do people who receive the greatest benefit from reliable automation also experience the greatest cost when it makes an error?' This question was examined by setting automation reliability at 70% for acceptances and hand-offs, resulting in 30% of aircraft 'missed' by the automation (not changing colour and flashing when required action). The automation error events were randomly distributed after the first five minutes. Conflict detection remained 100% reliable, allowing a further question of whether there was when automation on one task fails there is a transfer of cost to another automation tasks which remain reliable. Situation awareness and workload were assessed as previously to allow comparison and determine the extent of automation cost to these. The findings of this thesis further our understanding of the impact of information processing automation applied in a complex real-world simulated environment and the interactions with cognition.

References

Adamcyzk, P. D., & Bailey, B. P. (2004). If not now, when?: The effects of interruption at different moments within task execution. *In Human Factors in Computing Systems: Proceedings of CHI'04* (pp. 271-278). New York: ACM Press

Allport, A., & Wylie, G. (2000). Task switching, stimulus-response bindings, and negative priming. *Control of cognitive processes: Attention and performance XVIII*, 35-70.

Anderson, J. R., & Lebiere, C. (1998). *Atomic components of thought*. Mahwah, NJ: Erlbaum.

Annett, J. (2002). Subjective rating scales: science or art? *Ergonomics*, 45(14), 966-987.

Arnell, K. M., & Jolicoeur, P. (1999). The attentional blink across stimulus modalities: Evidence for central processing limitations. *Journal of Experimental Psychology: Human Perception and Performance*, 25(3), 630.

Bacon, L. P., & Strybel, T. Z. (2013). Assessment of the validity and intrusiveness of online-probe questions for situation awareness in a simulated air-traffic-management task with student air-traffic controllers. *Safety science*, 56, 89-95.

Bakdash, J. Z., Marusich, L. R., Cox, K., Geuss, M. N., & Zaroukian, E. G. (2020). *The validity of situation awareness for performance: A meta-analysis*.

Balfe, N., Sharples, S., & Wilson, J. R. (2015). Impact of automation: Measurement of performance, workload and behaviour in a complex control environment. *Applied Ergonomics*, *47*, 52–64.

Barron, L. G., & Rose, M. R. (2017). Multitasking as a Predictor of Pilot Performance: Validity Beyond Serial Single-Task Assessments. *Military Psychology*, *29*(4), 316–326.

Battiste, V., & Bortolussi, M. R. (1988). *Assessment of pilot workload with the introduction of an airborne threat-alert system* (No. 881385). SAE Technical Paper.

Bedny, G., & Meister, D. (1999). Theory of activity and situation awareness. *International Journal of cognitive ergonomics*, 3(1), 63-72.

Beer, J. M., Fisk, A. D., & Rogers, W. A. (2014). Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction. *Journal of Human-Robot Interaction*, *3*(2), 74.

Bernhardt, K. A., Salomon, K. A., Ferraro, F. R., Crockett, R. E. J., Terrell, H. K., Petros, T., & Vacek, J. J. (2016). Individual differences in dynamic multitasking performance. *Proceedings of the Human Factors and Ergonomics Society*, 1254–1258.

Bliss, J. P., & Dunn, M. C. (2000). Behavioural implications of alarm mistrust as a function of task workload. *Ergonomics*, 43(9), 1283-1300.

Boeing (2009). *Statistical summary of commercial jet airplane accidents worldwide operations*. Retrieved from https://www.boeing.com › about_bca › pdf › statsum

Bowden, V. K., Griffiths, N., Strickland, L., & Loft, S. (2021). Detecting a Single Automation Failure: The Impact of Expected (But Not Experienced) Automation Reliability. *Human Factors*.

Brandenburger, N., Naumann, A., & Jipp, M. (2019). Task-induced fatigue when implementing high grades of railway automation. *Cognition, Technology and Work, 0123456789.*

Brandimonte, M. A., Einstein, G. O., & McDaniel, M. A. (Eds.) (1996). *Prospective memory. Theory and applications*. Hillsdale, NJ: Erlbaum

Broadbent, D. E., & Broadbent, M. H. P. (1987). From detection to identification: Response to multiple targets in rapid serial visual presentation. *Perception & Psychophysics*, 42,105-113.

Bühner, M., König, C. J., Pick, M., & Krumm, S. (2006). Working Memory Dimensions as Differential Predictors of the Speed and Error Aspect of Multitasking Performance. *Human Performance*, *19*(3), 253–275.

Burns, C. M. (2018). Automation and the Human Factors Race to Catch Up. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 83–85.

Cak, S., Say, B., & Misirlisoy, M. (2020). Effects of working memory, attention, and expertise on pilots' situation awareness. *Cognition, Technology and Work*, *22*(1), 85–94.

Calhoun, G., Draper, M., & Ruff, H. A. (2009). Effect of level of automation on unmanned aerial vehicle routing task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 197–201.

Calhoun, G., Gloria, L., Ruff, H. A., Heath, A., Draper, Mark, H., Wright, & Evan, J. (2011). Automation-Level Transference Effects in Simulated Multiple Unmanned Aerial Vehicle Control. *Journal of Cognitive Engineering and Decision Making*, *5*(1), 55–82.

Chancey, E. T., Brill, J. C., & Bliss, J. P. (2017). *Multitasking Environment: Are False Alarms Really Worse than Misses? 2007*, 1621–1625.

Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied ergonomics*, 74, 221-232.

Chen, J. Y. C., & Joyner, C. T. (2009). Concurrent Performance of Gunner's and Robotics Operator's Tasks in a Multitasking Environment. *Military Psychology (Taylor & Francis Ltd)*, *21*(1), 98–113.

Chen, J. Y. C., & Terrence, P. I. (2008). Effects of tactile cueing on concurrent performance of military and robotics tasks in a simulated multitasking environment. *Ergonomics*, 51(8), 1137-1152.

Chen, J. Y. C., & Terrence, P. I. (2009). Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment. *Ergonomics*, *52*(8), 907–920.

Chen, S. I., Visser, T. A. W., Huf, S., & Loft, S. (2017). Optimizing the balance between task automation and human manual control in simulated submarine track management. *Journal of Experimental Psychology: Applied*, *23*(3), 240–262.

Chérif, L., Wood, V., Marois, A., Labonté, K., & Vachon, F. (2018). Multitasking in the military: Cognitive consequences and potential solutions. *Applied Cognitive Psychology*, *32*(4), 429–439.

Chiappe, D., Strybel, T. Z., & Vu, K. P. L. (2015). A situated approach to the understanding of dynamic situations. *Journal of Cognitive Engineering and Decision Making*, 9(1), 33-43.

Chiappe, D., Vu, K. P. L., Rorie, C., & Morgan, C. (2012). A situated approach to shared situation awareness. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56,(1), 748-752. Sage CA: Los Angeles, CA: SAGE Publications.

Chun, M. M., & Potter, M. C. (2001). The attentional blink and task switching within and across modalities. *The limits of attention: Temporal constraints in human information process*sing, 20-35.

Comstock, J. R., & Arnegard, R. J. (1992). *The multi-attribute task battery for human operator workload and strategic behavior research*.

Cullen, R. H., Dan, C. S., Rogers, W. A., & Fisk, A. D. (2014). The Effects of Experience and Strategy on Visual Attention Allocation in an Automated Multiple-Task Environment. *International Journal of Human-Computer Interaction*, 30(7), 533-546.

Cullen, R. H., Rogers, W. A., & Fisk, A. D. (2013). Human performance in a multiple-task environment: Effects of automation reliability on visual attention allocation. *Applied Ergonomics*, *44*(6), 962–968.

De Jong, R. (2000). An Intention-Activation Account of Residual Switch Costs. *Control of cognitive processes*, 357.

Dell'Acqua, R., Jolicœur, P., Luria, R., & Pluchino, P. (2009). Reevaluating encoding-capacity limitations as a cause of the attentional blink. *Journal of Experimental Psychology: Human Perception and Performance*, 35(2), 338.

Deng, Y., Shirley, J., Zhang, W., Kim, N. Y., & Kaber, D. (2020). Influence of dynamic automation function allocations on operator situation awareness and workload in unmanned aerial vehicle control. In *Advances in Intelligent Systems and Computing*, (959). Springer International Publishing.

Di Nocera, F., Fabrizi, R., Terenzi, M., & Ferlazzo, F. (2006). Procedural errors in air traffic control: Effects of traffic density, expertise, and automation. In *Aviation, space, and environmental medicine*, 7.

Di Nocera, F., Lorenz, B., & Parasuraman, R. (2005). Consequences of shifting from one level of automation to another: main effects and their stability. *Human Factors in Design, Safety, and Management*, 2005, 363–376.

Dikmen, M., & Burns, C. (2016). Autonomous Driving in the Real World: Experiences with Tesla Autopilot and Summon. *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '16), Ann Arbor, MI, USA., May*, 225–228.

Dodhia, R. M., & Dismukes, R. K. (2005). A task interrupted becomes a prospective memory task. *In Biennial Meeting of the Society for Applied Research in Memory and Cognition, Wellington, New Zealand*.

Duncan, J., Martens, S., & Ward, R. (1997). Restricted attentional capacity within but not between sensory modalities. *Nature*, 387(6635), 808-810.

Durso, F. T., Dattel, A. R., Banbury, S., & Tremblay, S. (2004). SPAM: The real-time assessment of SA. *A cognitive approach to situation awareness: Theory and application*, 1, 137-154.

Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., Nikolic, D., & Manning, C. A. (1998). Situation Awareness As a Predictor of Performance in En Route Air Traffic Controllers. *Air Traffic Control Quarterly*, *6*, 1–20.

Durso, F. T., & Sethumadhavan, A. (2008). Situation awareness: Understanding dynamic environments. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(3), 442–448.

Dzindolet, M. T., Pierce, L., Pomranky, R., Peterson, S., & Beck, H. (2001). Automation Reliance On A Combat Identification System. *Proceedings Of The Human Factors And Ergonomics Society*, *45*(4), 532–536.

Edwards, T., Homola, J., Mercer, J., & Claudatos, L. (2017). Multifactor interactions and the air traffic controller: the interaction of situation awareness and workload in association with automation. *Cognition, Technology and Work*, 19(4), 687–698.

Elsmore, T. F. (1994). SYNWORK1: A PC-based tool for assessment of performance in a simulated work environment. *Behavior Research Methods, Instruments, & Computers*, 26(4), 421-426.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. *In Proceedings of the Human Factors Society annual meeting*, 32(2), 97-101.

Endsley, M. R. (1995). Toward a Theory of Situation Awareness. *Human Factors*, *37*(1), 32–64.

Endsley, M. R. (2000). Theoretical Underpinnings of Situation Awareness: A Critical Review Process, 3-32

Endsley, M. R. (2015). Situation Awareness Misconceptions and Misunderstandings. *Journal of Cognitive Engineering and Decision Making*, 9(1), 4–32.

Endsley, M. R. (2017a). From Here to Autonomy: Lessons Learned from Human-Automation Research. *Human Factors*, 59(1), 5–27.

Endsley, M. R. (2017b). Autonomous Driving Systems: A Preliminary Naturalistic Study of the Tesla Model S. *Journal of Cognitive Engineering and Decision Making*, 11(3), 225–238.

Endsley, M. R. (2019). A Systematic Review and Meta-Analysis of Direct Objective Measures of Situation Awareness: A Comparison of SAGAT and SPAM. *Human Factors*, 63(1), 124-150.

Endsley, M. R. (2020). The divergence of objective and subjective situation awareness: A meta-analysis. *Journal of cognitive engineering and decision making*, 14(1), 34-53.

Endsley, M., & Jones, W. M. (1997). *Situation Awareness Information Dominance & Information Warfare.* Logicon Technical Services INC, Dayton OH.

Endsley, M. R., & Kaber, D. B. (1997). The use of level of automation as a means of alleviating out-of-the-loop performance problems: A taxonomy and empirical analysis. In *13th Triennial Congress of the International Ergonomics Association*, (1), 168-170. Finnish Institute of Occupational Health Helsinki.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human factors*, 37(2), 381-394.

Falkland, E. C., & Wiggins, M. W. (2019). Cross-task cue utilisation and situational awareness in simulated air traffic control. *Applied Ergonomics*, 74(January 2018), 24–30.

Fitts, P. M. (1951). *Human engineering for an effective air-navigation and traffic-control system*.

Fothergill, S., Loft, S., & Neal, A. (2009). ATC–labAdvanced: An air traffic control simulator with realism and control. *Behavior Research Methods*, 41(1), 118-127.

Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007). Measurement of trust in human-robot collaboration. In *2007 International symposium on collaborative technologies and systems*, 106-114. Institute of Electrical and Electronics Engineers.

Galster, S. M., Bolia, R. S., & Parasuraman, R. (2002). Effects of information automation and decision-aiding cueing on action implementation in a visual search task. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 46(3), 438-442. Sage CA: Los Angeles, CA: SAGE Publications.

Gawron, V. J. (2008). *Human performance, workload, and situational awareness measures handbook*. Crc Press.

Gutzwiller, R. S., Wickens, C. D., & Clegg, B. A. (2019). The role of reward and effort over time in task switching. *Theoretical Issues in Ergonomics Science*, 20(2), 196–214.

Hambrick, D. Z., Oswald, F. L., Darowski, E. S., Rench, T. A., & Brou, R. (2010). Predictors of multitasking performance in a synthetic work paradigm. *Applied Cognitive Psychology*, 24(8), 1149–1167.

Hancock, P. A., Jagacinski, R. J., Parasuraman, R., Wickens, C. D., Wilson, G. F., & Kaber, D. B. (2013). Human-Automation Interaction Research. *Ergonomics in Design: The Quarterly of Human Factors Applications*, 21(2), 9–14.

Harris, W., Hancock, P. A., Arthur, E., & Caird, J. K. (1995). *Performance, Workload, and Fatigue Changes Associated with Automation*, 169–185.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *In Advances in psychology*, 52, 139-183. North-Holland.

Hart, S. G., & Wickens, C. D. (2010). Cognitive workload. *NASA human systems integration handbook*, 1-17.

Helmreich, R. L., & Foushee, H. C. (1993). *Why crew resource management? Empirical and theoretical bases of human factors training in aviation*. Academic Press.

Hopkin, V.D., (1995). *Human Factors in Air Traffic Control* (London: Taylor & Francis).

Inners, M., & Kun, A. L. (2017, September). Beyond liability: Legal issues of human-machine interaction for automated vehicles. *In Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 245-253).

Jipp, M., & Ackerman, P. L. (2016). The Impact of Higher Levels of Automation on Performance and Situation Awareness. *Journal of Cognitive Engineering and Decision Making,* 10(2), 138–166.

John, M. S., & Smallman, H. S. (2008). Staying up to speed: Four design principles for maintaining and recovering situation awareness. *Journal of Cognitive Engineering and Decision Making*, 2(2), 118-139.

Jolicœur, P., & Dell'Acqua, R. (1999). Attentional and structural constraints on visual encoding. *Psychological research*, *62*(2), 154-164.

Jolicœur, P., Dell'Acqua, R., & Crebolder, J. M. (2001). The attentional blink bottleneck. In K. *Shapiro (Ed.), The limits of attention: Temporal constraints in human information processing* (pp. 82–99). Oxford University Press

Kaber, D. B. (2018). Issues in Human–Automation Interaction Modeling: Presumptive Aspects of Frameworks of Types and Levels of Automation. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 7–24

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*(2), 113–153.

Kaber, D. B., Perry, C. M., Segall, N., McClernon, C. K., & Prinzel, L. J. (2006). Situation awareness implications of adaptive automation for information processing in an air traffic control-related task. *International Journal of Industrial Ergonomics*, *36*(5), 447–462.

Keeler, J., Battiste, H., Hallett, E. C., Roberts, Z., Winter, A., Sanchez, K., Strybel, T. Z., & Vu, K.-P. L. (2015). May I Interrupt? The effect of SPAM Probe Questions on Air Traffic Controller Performance. *Procedia Manufacturing*, *3*, 2998–3004.

Konig, C. J., Buhner, M., & Murling, G. (2005). Working memory, fluid intelligence, and attention are predictors of multitasking performance, but polychronicity and extraversion are not. *Human performance*, *18*(3), 243-266.

Körber, M., Weißgerber, T., Kalb, L., Blaschke, C., & Farid, M. (2015). *Prediction of take-over time in highly automated driving by two psychometric tests*. *82*(193), 195–201.

Lau, S. K. (2007). *Fly with intelligence – best practices to improve safety and efficiency of flight operations.* CAPACG (white paper).

Li, Y., & Burns, C. M. (2017). Modeling Automation With Cognitive Work Analysis to Support Human-Automation Coordination. *Journal of Cognitive Engineering and Decision Making*, *11*(4), 299–322.

Lin, J., Matthews, G., Wohleber, R., Chiu, C. Y. P., Calhoun, G., Funke, G., & Ruff, H. (2016). Automation reliability and other contextual factors in multi-UAV operator selection. *Proceedings of the Human Factors and Ergonomics Society*, *2015*, 845–849.

Loft, S., Bowden, V., Braithwaite, J., Morrell, D. B., Huf, S., & Durso, F. T. (2015). Situation awareness measures for simulated submarine track management. *Human factors*, 57(2), 298-310.

Loft, S., Chapman, M., & Smith, R. E. (2016). Reducing prospective memory error and costs in simulated air traffic control: External aids, extending practice, and removing perceived memory requirements. *Journal of experimental psychology: applied*, 22(3), 272.

Loft, S., Sadler, A., Braithwaite, J., & Huf, S. (2015). The chronic detrimental impact of interruptions in a simulated submarine track management task. *Human factors*, 57(8), 1417-1426.

Longo, L. (2018). Experienced mental workload, perception of usability, their interaction and impact on task performance. *Public Library of Science one*, 13(8), e0199661.

Lyons, J. B., Ho, N. T., Fergueson, W. E., Sadler, G. G., Cals, S. D., Richardson, C. E., & Wilkins, M. A. (2016). Trust of an automatic ground collision avoidance technology: A fighter pilot perspective. *Military Psychology*, *28*(4), 271–277.

Manzey, D., Luz, M., Mueller, S., Dietz, A., Meixensberger, J., & Strauss, G. (2011). Automation in surgery: The impact of navigated-control assistance on performance, workload, situation awareness, and acquisition of surgical skills. *Human Factors*, *53*(6), 584–599.

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87

Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in cognitive sciences*, 9(6), 296-305.

Masakowski, Y. R., & Creely, T. E. (2017). Leaders and Ethical Decision Making with Autonomous Systems. *Naturalistic Decision Making and Uncertainty*, 125.

Mansikka, H., Virtanen, K., & Harris, D. (2018). Dissociation between mental workload, performance, and task awareness in pilots of high performance aircraft. *IEEE Transactions on Human-Machine Systems*, 49(1), 1-9.

Matthews, M. D., & Beal, S. A. (2002). Assessing situation awareness in field training exercises. *Military Academy West Point NY Office of Military Psychology and Leadership*.

Matthews, G., De Winter, J., & Hancock, P. A. (2020). What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. *Theoretical issues in ergonomics science*, 21(4), 369-396.

Matthews, M. D., Shattuck, L. G., Graham, S. E., Weeks, J. L., Endsley, M. R., & Strater, L. D. (2001). Situation Awareness for Military Ground Forces: Current Issues and Perspectives. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *45*(4), 351–355.

McGarry, K., Rovira, E., & Parasuraman, R. (2003). Effects of task duration and type of automation support on human performance and stress in a simulated battlefield engagement task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *47*(3), 548–552.

Meiran, N., Chorev, Z., & Sapir, A. (2000). Component processes in task switching. *Cognitive psychology*, 41(3), 211-253.

Monk, C. A., Trafton, J. G., & Boehm-Davis, D. A. (2008). The effect of interruption duration and demand on resuming suspended goals. *Journal of experimental psychology: Applied*, 14(4), 299.

Monsell, S. (2003). Task switching. *Trends in cognitive sciences*, 7(3), 134-140.

Moray, N., (1959) Attention in dichotic listening: affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, 11:56–60.

Moray, N., (1979). *Mental Workload: Its Theory and Measurement* (New York: Plenum).

Morgan, B., D'Mello, S., Abbott, R., Radvansky, G., Haass, M., & Tamplin, A. (2013). Individual differences in multitasking ability and adaptability. *Human Factors*, *55*(4), 776–788.

Muslim, H., & Itoh, M. (2019). A theoretical framework for designing human-centered automotive automation systems. *Cognition, Technology and Work*, *21*(4), 685–697.

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological review*, 86(3), 214.

Nguyen, T., Lim, C. P., Nguyen, N. D., Gordon-Brown, L., & Nahavandi, S. (2019). A Review of Situation Awareness Assessment Approaches in Aviation Environments. *IEEE Systems Journal*, *13*(3), 3590–3603.

Ockerman, J. J., & Pritchett, A. R. (2002). Impact of Contextual Information on Automation Brittleness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 46(3), 382–386.

O'Brien, K. S., & O'Hare, D. (2007). Situational awareness ability and cognitive skills training in a complex real-world task. *Ergonomics*, *50*(7), 1064–1091.

Onnasch, L., Ruff, S., & Manzey, D. (2014). Operators′ adaptation to imperfect automation – Impact of miss-prone alarm systems on attention allocation and performance. *International Journal of Human-Computer Studies*, *72*(10–11), 772–782.

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *56*(3), 476–488.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions On*, *30*(3), 286–297.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of Cognitive Engineering and Decision Making*, *2*(2), 140–160.

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still Vital After All These Years of Automation. *Human Factors*, *50*(3), 511–520.

Pashler, H. (1984). Processing stages in overlapping tasks: Evidence for a central bottleneck. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(3), 358–377.

Pashler, H. (1992). Attentional Limitations in Doing Two Tasks at the Same Time. *Current Directions in Psychological Science*, *1*(2), 44–48.

Pickering, M. J., & Garrod, S. (2004). The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, 27(2), 212-225.

Pritchett, A. R., Kim, S. Y., & Feigh, K. M. (2014). Modeling human–automation function allocation. *Journal of cognitive engineering and decision making*, 8(1), 33-51.

Rafferty, L., Ladva, D., Salmon, P. M., Young, M., Jenkins, D., Walker, G. H., & Stanton, N. A. (2008). Measuring Situation Awareness in complex systems: Comparison of measures study. *International Journal of Industrial Ergonomics*, *39*(3), 490–500.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink?. *Journal of experimental psychology: Human perception and performance*, 18(3), 849.

Ratwani, R. M., Trafton, J. G., & Myers, C. (2006). Helpful or harmful? Examining the effects of interruptions on task performance. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(3), 372-375. Sage CA: Los Angeles, CA: SAGE Publications.

Ratwani, R., & Trafton, J. G. (2010). An eye movement analysis of the effect of interruption modality on primary task resumption. *Human factors*, 52(3), 370-380.

Redick, B. T. S., Shipstead, Z., Meier, M. E., Montroy, J. J., Hicks, K. L., Unsworth, N., Kane, M. J., & Hambrick, D. Z. (2016). Cognitive predictors of a common multitasking ability: Contributions from working memory, attention control, and fluid intelligence. *Journal of experimental psychology: General*, 145(11), 1473.

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. *In Advances in psychology*, 52, 185-218. North-Holland.

Riley, V. (1996). What avionics engineers should know about pilots and automation. IEEE *Aerospace and Electronic Systems Magazine*, 11(5), 3-8.

Roscoe, A. H., & Ellis, G. A. (1990). A subjective rating scale for assessing pilot workload in flight: A decade of practical use. *Royal Aerospace Establishment Farnborough (United Kingdom)*.

Roth, E. M., Sushereba, C., Militello, L. G., Diiulio, J., & Ernst, K. (2019). Function Allocation Considerations in the Era of Human Autonomy Teaming. *Journal of Cognitive Engineering and Decision Making*, *13*(4), 199–220.

Rouse, W. B. (1988). Adaptive aiding for human/computer control. *Human Factors*, 30(4), 431-443.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(1), 76–87.

Rovira, E., Zinni, M., & Parasuraman, R. (2002). *On Multi-task Performance*. *2000*, 327–331.

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied psychology*, 53(1), 61-86.

Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive Control of Cognitive Processes in Task Switching. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(4), 763–797.

Salmon, P. M., Stanton, N. A., Walker, G. H., Jenkins, D., Ladva, D., Rafferty, L., & Young, M. (2009). Measuring Situation Awareness in complex systems: Comparison of measures study. *International Journal of Industrial Ergonomics*, *39*(3), 490–500.

Stanton, N. A. (2016). Distributed situation awareness. Theoretical Issues in Ergonomics Science, 17(1), 1-7.

Stanton, N. A., Hedge, A., Brookhuis, K., Salas, E., & Hendrick, H. W. (Eds.). (2004). *Handbook of human factors and ergonomics methods.* CRC press.

Stanton, N. A., Salmon, P. M., Walker, G. H., & Jenkins, D. (2009). Genotype and phenotype schemata and their role in distributed situation awareness in collaborative systems. *Theoretical Issues in Ergonomics Science*, 10(1), 43-68.

Stanton, N. A., Stewart, R., Harris, D., Houghton, R. J., Baber, C., McMaster, R., ... & Green, D. (2006). Distributed situation awareness in dynamic systems: theoretical development and application of an ergonomics methodology. *Ergonomics*, 49(12-13), 1288-1311.

Sanchez, J. (2009). Conceptual Model of Human-Automation Interaction. *Proceedings of the Human Factors and Ergonomics Society*, 798–803.

Santiago-Espada, Y., Myer, R. R., Latorella, K. A., & Comstock Jr, J. R. (2011). *The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide.*

Saqer, H., & Parasuraman, R. (2014). Individual performance markers and working memory predict supervisory control proficiency and effective use of adaptive automation. *International Journal of Human Factors and Ergonomics*, *3*(1), 15–31.

Sarter, N. B., & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors*, 37(1), 5-19.

Sauer, J., Kao, C.-S., & Wastell, D. (2012). A comparison of adaptive and adaptable automation under different levels of environmental stress. *Ergonomics*, *55*(8), 840–853.

Sauer J, Wastell DG, Hockey GR, Earle F. (2003). Performance in a complex multiple-task environment during a laboratory-based simulation of occasional night work. *Human Factors.*,45(4), 657-69.

Saxby, D. J., Matthews, G., Hitchcock, E. M., Warm, J. S., Funke, G. J., & Gantzer, T. (2008, September). Effect of active and passive fatigue on performance using a driving simulator. In *Proceedings of the Human Factors and Ergonomics Society Annual Meetin*, 52(21), 1751-1755. Sage CA: Los Angeles, CA: Sage Publications.

Schumacher, E. H., Seymour, T. L., Glass, J. M., Fencsik, D. E., Lauber, E. J., Kieras, D. E., & Meyer, D. E. (2001). Virtually Perfect Time Sharing in Dual-Task Performance: Uncorking the Central Cognitive Bottleneck. *Psychological Science*, *12*(2), 101–108.

Sebok, A., & Wickens, C. D. (2017). Implementing Lumberjacks and Black Swans into Model-Based Tools to Support Human-Automation Interaction. *Human Factors*, *59*(2), 189–203.

Sethumadhavan, A. (2009). Effects Of Automation Types on Air Traffic Controller Situation Awareness and Performance. *Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting—2009 1 Effects*, 1329–1333.

Sheridan, T. B., & Verplank, W. L. (1978). Human and Computer Control of Undersea Teleoperators. *ManMachine Systems Lab Department of Mechanical Engineering MIT Grant N0001477C0256, DECEMBER 1978*, 343.

Smith, K., & Hancock, P. A. (1995). Situation awareness is adaptive, externally directed consciousness. *Human Factors*, 37(1), 137-148.

Sorensen, L. J., & Stanton, N. A. (2016). Inter-rater reliability and content validity of network analysis as a method for measuring distributed situation awareness. *Theoretical Issues in Ergonomics Science*, *17*(1), 42-63.

Stein, E. S. (1998). Human operator workload in air traffic control. *Human factors in air traffic control*.

Strybel, T. Z., Keeler, J., Mattoon, N., Alvarez, A., Barakezyan, V., Barraza, E., ... & Battiste, V. (2017, July). Measuring the effectiveness of human autonomy teaming. *In International Conference on Applied Human Factors and Ergonomics*, 23-33. Springer, Cham.

Strybel, T. Z., Vu, K. P. L., Chiappe, D. L., Morgan, C. A., Morales, G., & Battiste, V. (2016). Effects of NextGen concepts of operation for separation assurance and interval management on Air Traffic Controller situation awareness, workload, and performance. *The International Journal of Aviation Psychology*, *26*(1-2), 1-14.

Squire, P. N., Galster, S. M., & Parasuraman, R. (2004). The effects of levels of automation in the human control of multiple robots in the RoboFlag simulation environment. *Human performance, situation awareness, and automation: Current research and trends*, 2, 48-53.

Squire, P. N., & Parasuraman, R. (2010). Effects of automation and task load on task switching during human supervision of multiple semi-autonomous robots in a dynamic environment. *Ergonomics*, *53*(8), 951–961.

Strand, N., Nilsson, J., Karlsson, I. C. M., & Nilsson, L. (2014). Semi-automated versus highly automated driving in critical situations caused by automation failures. *Transportation Research Part F: Traffic Psychology and Behaviour*, *27*, 218–228.

Sudevan, P., & Taylor, D. A. (1987). The cuing and priming of cognitive operations. *Journal of Experimental Psychology: Human perception and performance*, 13(1), 89.

Tatasciore, M., Bowden, V. K., Visser, T. A. W., Michailovs, S. I. C., & Loft, S. (2019). The Benefits and Costs of Low and High Degree of Automation. *Human Factors*.

Taylor, R. M., & Selcon, S. J. (1990, October). Cognitive quality and situational awareness with advanced aircraft attitude displays. *In Proceedings of the Human Factors Society Annual Meeting*, 34(1), 26-30. Sage CA: Los Angeles, CA: SAGE Publications.

Thomas, M. L., & Russo, M. B. (2007). Neurocognitive monitors: Towards the prevention of cognitive performance decrements and catastrophic failures in the operational environment. *Aviation, Space and Environ- mental Medicine*, 78(5 Suppl), B144–B152

Tombu, M., & Jolicœur, P. (2003). A central capacity sharing model of dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance*, 29(1), 3–18.

Tran Luciani, D., Löwgren, J., & Lundberg, J. (2019). Designing fine-grained interactions for automation in air traffic control. *Cognition, Technology and Work*.

Trafton, J. G., Altmann, E. M., Brock, D. P., & Mintz, F. E. (2003). Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal., *International Journal of Human-Computer Studies*, 58(5), 583-603.

Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, 12(4), 242-248.

Tsang, P. S., & Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, *39*(3), 358-381.

Vidulich, M. A., & Hughes, E. R. (1991, September). Testing a subjective metric of situation awareness. *In Proceedings of the Human Factors Society Annual Meeting*, 35(18), 1307-1311. Sage CA: Los Angeles, CA: SAGE Publications.

Vidulich, M. A., & Tsang, P. S. (2012). Mental workload and situation awareness.(chap. 8), and Salvendy, G.(Ed.). *Handbook of Human Factors and Ergonomics*.

Vu, K. L., Strybel, T. Z., Battiste, V., Lachter, J., Dao, A. V., Brandt, S., Ligda, S., & Johnson, W. (2012). Pilot performance in trajectory-based operations under concepts of operation that vary separation responsibility across pilots, air traffic controllers, and automation. *International Journal of Human–Computer Interaction*, 28, 107–118

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance Requires Hard Mental Work and Is Stressful. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 433–441.

Welford, A. T. (1952). The 'psychological refractory period' and the timing of high-speed performance—a review and a theory. *British Journal of Psychology*. General Section, 43(1), 2-19.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, 3(2), 159-177.

Wickens, C. D., & Boles, D. B. (1983). The limits of multiple resource theory: The role of task correlation/integration in optimal display formatting. *Illinois University at Urbana Engineering-Psychology Research Lab.*

Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and Automation Bias in the Use of Imperfect Automation. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *57*(5), 728–739.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: a synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212.

Wickens, C. D., Gutzwiller, R. S., & Santamaria, A. (2015). Discrete task switching in overload: A meta-analyses and a model. *International Journal of Human Computer Studies*, *79*, 79–84

Wickens, C. D., Gutzwiller, R. S., Vieane, A., Clegg, B. A., Sebok, A., & Janes, J. (2016). Time Sharing between Robotics and Process Control: Validating a Model of Attention Switching. *Human Factors*, *58*(2), 322–343.

Wickens, C.D., & Hollands, J.G. (2000). *Engineering Psychology and Human Performance*. *Upper Saddle River*, NJ: Prentice Hall

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. B. (2010). Stages and Levels of Automation: An Integrated Meta-analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(4), 389–393.

Wickens, C. D., McCarley, J. S., Alexander, A. L., Thomas, L. C., Ambinder, M., Zheng, S., & Field, M. (2005). Model of Pilot Error. *Contract*, *January*, 213.

Wickens, C. D., Santamaria, A., & Sebok, A. (2013, September). A computational model of task overload management and task switching. In *Proceedings of the human factors and ergonomics society annual meeting*, 57(1), 763-767. Sage CA: Los Angeles, CA: SAGE Publications.

Wood, N., Cowan, N. (1995) The cocktail party phenomenon revisited: how frequent are attention shifts to one's name in an irrelevant auditory channel? *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 21:255–260.

Woodhouse, R., & Woodhouse, R.A. (1995). Navigation errors in relation to controlled flight into terrain (CFIT) accidents. *In Proceedings of the 8th International Symposium on Aviation Psychology*, R. Jensen (Ed.), Columbus, OH: Ohio State University, (pp 45–50).

Wong, C. Y., & Seet, G. (2017). Workload, awareness and automation in multiple-robot supervision. *International Journal of Advanced Robotic Systems*, *14*(3), 1–16.

Wright, J. L., Chen, J. Y. C., & Barnes, M. J. (2018). Human-Automation Interaction for Multiple Robot Control: The Effect of Varying Automation Assistance and Individual Differences on Operator Performance. *Ergonomics*, 1–34.

Yeh, Y. Y., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human factors*, 30(1), 111-120.

Zijlstra, F.R.H., 1993. Efficiency in work behavior. a design approach for modern tools. *PhD thesis, Delft University of Technology*. Delft, The Netherlands: Delft University Press. Book

CHAPTER 2:

Individual differences in multi-tasking ability modulate the benefits of using low-degree automation

Chapter Abstract

Background: One conventional approach to augment human performance and reduce workload is to provide operators with low-degree automation (DOA). However, no prior work has examined whether the impact of low-DOA differs across individuals as a function of their cognitive ability. **Methods:** One-hundred-and-six undergraduate students completed three cognitive tasks (Dual task, Psychological Refractory Period, Attentional Blink) which were combined to create a latent factor indexing multi-tasking ability. Participants completed two conditions of a simulated air traffic control task: once with no automation (manual) and once with low-DOA to assist aircraft acceptance and hand-off tasks. Conflict detection was performed without automation in both conditions. **Results:** Compared to performance in the manual condition, individuals with poorer multi-tasking ability benefited more from automation on the aircraft acceptance and hand-off tasks than individuals with higher multi-tasking ability. Higher multi-tasking ability also led to better conflict detection performance, situation awareness and objective workload but did not modulate the benefit of automation on situation awareness or subjective workload. Participants with poorer multi-tasking ability performed more poorly and had lower SA but benefited more on tasks supported by low-DOA than individuals with greater multi-tasking ability. **Conclusions:** Low-DOA can lead to greater performance benefits when used to support individuals with low multi-tasking ability, suggesting the need to consider individual differences in cognitive ability when designing and implementing automation in the workplace.

Introduction

Humans are frequently required to perform concurrent tasks under time pressure to respond safely and efficiently in complex work domains, including patient monitoring (Wickens, Clegg et al., 2015), driving (Körber et al., 2015), and ATC (Sethumadhavan, 2009). However, when task demands exceed humans' cognitive capacity, workload can increase and performance can decline (Wickens, et al., 2010). To combat this problem, automation has been deployed to assist operators to free-up sufficient cognitive resources to avoid overwhelming their capacity (Kaber & Endsley, 2004; Saqer & Parasuraman, 2014), thereby reducing sub-optimal performance and serious accidents (Nguyen et al., 2019; Thomas & Russo, 2007).

The impact of automation is often assessed in terms of three inter-related outcomes: performance, workload, and situation awareness (SA; Kaber & Endsley, 2004; Wright et al., 2018; see Onnasch et al., 2014 for a meta-analysis of studies involving these outcomes). Performance focuses on metrics such as task accuracy and response time (Manzey et al., 2012). Workload focuses on the mental or physical activity required of the human operator for successful performance (Parasuraman et al., 2000) and can be measured subjectively (e.g., self-report) or objectively (e.g., secondary tasks). Finally, SA focuses on "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1988b, p. 97), and can be measured subjectively by ratings scales (e.g., Situation Awareness Rating Technique; Taylor & Selcon, 1990), or objectively by queries (e.g., Situation Awareness Global Assessment Technique; Endsley, 1995; Situation Present Assessment Method; Durso et al., 2004) or observer ratings (Endsley, 2019b).

Research comparing these outcomes across conditions (i.e., with, and without automation, called 'manual') has typically found introducing automation leads to robust performance benefits (e.g., Manzey et al., 2012; Wright et al., 2018), and lower subjective and objective workload compared to manual conditions (Parasuraman et al., 2009; Strybel et al., 2016). However, evidence concerning SA has been more variable, with some studies finding automation-related benefits (e.g., Kaber & Endsley, 2004; Parasuraman et al., 2009), and others automation-related impairments (Strybel et al., 2016) compared to manual conditions.

Automation outcomes vary

Considerable research has also focused on how outcomes vary as a function of the *degree of automation* (DOA; Wickens et al., 2010). The DOA taxonomy is based on joint consideration of the level of control that automation exerts (ranging from no action to complete; Sheridan & Verplank, 1978) and the stage of human information processing that automation supplements (acquisition, analysis, decision selection/action recommendation, and action implementation;

Parasuraman et al., 2000). Research suggests that as DOA increases, performance improves (Endsley & Kaber, 1999; Manzey et al., 2012) and subjective workload decreases (Manzey et al., 2012; Rovira et al., 2007), but SA declines (Endsley & Kiris, 1995; Kaber, et al., 2000; for a summary see Onnasch et al., 2014).

Past work has focused on how variations in DOA modulate outcome metrics (see Wickens et al., 2010). However, a key difference is that whereas this earlier work looked at how individual differences modulated the impact of different DOAs within the same individual (e.g., Jipp & Ackerman, 2016; Wright et al., 2018), this chapter will examine how individual differences modulate the impact of the *same* DOA across *different* individuals, as well as *within* an individual's performance. Further, no previous studies I am aware of have investigated multi-tasking as a cognitive ability. Earlier studies have examined other individual differences, including spatial ability, information processing and working memory capacity (Jipp & Ackerman, 2016; Wright et al., 2018) and have suggested these can modulate the impact of increasing DOA. This possibility is consistent with the fact that automation is designed to augment human cognition. Thus, one would expect an operator with less cognitive ability should benefit proportionally more from automation designed to augment that cognitive ability than an operator with more cognitive ability. Such an outcome would be of considerable practical importance as it would imply that individuals could benefit differentially from the same automation. For example, low-DOA, which to-date has shown to be of less benefit to the operator than high-DOA (Onnasch et al., 2014), may be of considerably greater value than currently thought for operators with poorer cognitive abilities.

Multi-tasking ability

To test this novel prediction, in the present chapter participants completed a simulated ATC task that required them to *multi-task*: monitoring their airspace to "accept" incoming aircraft and "hand-off" outgoing aircraft, while performing "conflict detection" to avoid aircraft violating minimum separation standards (Fothergill et al., 2009). Multi-tasking can be broadly defined as "the strategic allocation of resources among multiple tasks" (Pashler, 1984, 1992). It should be noted that 'pure multi-tasking', the concept of concurrent attention and awareness of two sensory inputs from the same modality (i.e., auditory, or visual) is structurally limited and not possible (Duncan et al., 1997; Marois & Ivanoff, 2005). However, multi-tasking is commonly operationalised in terms three key component processes (Chérif et al., 2018): (1) performing tasks concurrently (dual-tasking), (2) alternating between tasks (task switching), and (3) task interruption. Crucially, multi-tasking ability is also widely variable in the general population (Redick et al., 2016), associated with stable individual differences in cognitive ability (working memory, Redick, et al., 2016; attentional control, J. Chen & Terrence, 2009; reasoning, Bühner et al., 2006), and strongly linked to performance in complex tasks (Wickens et al., 2015). Multi-

tasking as a cognitive ability has not been previously investigated with respect to performance with automation. Therefore, the reduction of task demands (i.e., multi-tasking requirements) is a central assumption underlying the implementation of automated systems (Kaber & Endsley, 2004; Saqer & Parasuraman, 2014).

The current study

The primary aim of this chapter was to investigate whether individual differences in multi-tasking ability modulated the benefit of perfectly reliable automation on performance, SA, and workload in an ATC task. To this end, participants completed one ATC condition manually and one with perfectly reliable automation. In the automation condition, low-DOA visually highlighted aircraft requiring acceptance or hand-off. Conflict detection was not subject to automation in this chapter. Computerised tasks that tapped into the three core elements of multi-tasking (Chérif et al., 2018) were used to assess individual differences. These were chosen over a tasked-based measure of multi-tasking (such as the Multi-Attribute Task Battery-II; Comstock & Arnegard, 1992) as the three tasks have previously been investigated together (Bender et al., 2018; Redick et al., 2016) and have been extensively used in previous literature to examine the cognitive costs of performing multiple tasks (e.g., Johnston & Pashler, 1998; Klapp et al., 2019; Tombu & Jolicoeur, 2005). Data from these tasks were concatenated using factor-analytic techniques to form a multi-tasking latent factor (Redick et al., 2016). Finally, to assess these outcomes, this chapter measured performance on the acceptance, hand-off, and conflict detection tasks. SA was assessed using the Situation Present Awareness Method (Durso et al., 2004) which was chosen for its previous use in ATC (Cak et al., 2020; Pierce et al., 2008) and benefits over other common SA measures in the current design (see Chapter 1 for discussion). Both subjective and objective measures of workload were used.

Several predictions were made with regards to the effects of automation and multi-tasking separately and in combination. In keeping with past literature, it was predicted low-DOA would benefit performance on the acceptance and hand-off tasks and reduce subjective and objective workload compared to the manual baseline. However, as noted above, previous studies have found inconsistent effects of automation on SA compared to manual conditions, and so this chapter made no prediction. With respect to multi-tasking ability, this chapter predicted that better multi-taskers would out-perform and have lower subjective and objective workload than poorer multi-taskers. This follows from literature linking increased workload and poorer performance to conditions in which task demands exceed operator cognitive capacity (Saqer & Parasuraman, 2014). Finally, given evidence for cognitive abilities, most notably working memory, are central to both multi-tasking and SA (Gutzwiller et al., 2013; Redick, 2016), it was also expected that better multi-taskers would demonstrate better SA than poorer multi-taskers.

With respect to the key area of interest – the moderating impact of multi-tasking ability on automation – it was expected automation benefits for performance, and subjective and objective workload would be largest for poorer multi-taskers and decline in strength as multi-tasking ability increased. This is in keeping with past literature which argued that automation reduces cognitive demands on operators, and thereby improves outcomes. At first glance, this literature would also seem to suggest that SA for poorer multi-taskers should benefit more from automation as well. However, this expectation is tempered by the inconsistent evidence for automation benefits on SA. If automation does not benefit SA, which would imply that it is not aiding SA-related cognitions, then it is also less likely that automation would interact with multi-tasking ability.

Finally, the present design also allowed this chapter to investigate the impact of multi-tasking ability and automation on the non-automated conflict detection task. There is a clear expectation that better multi-taskers should outperform poorer multi-taskers on this task. However, the influence of automating the acceptance and hand-off tasks on the non-automated conflict detection task is less clear. While some studies suggested automation leads to performance benefits on unrelated concurrent tasks (e.g., Kaber & Endsley, 2004), others have shown decrements when the automated and unautomated tasks depend on processing common display information (S. Chen et al., 2017; Tatasciore, et al., 2020; Tatasciore, et al., 2021). Given that conflict detection depends on display information initially used during the acceptance task (i.e., tracking aircraft as they approach the sector and following their acceptance into the sector), it was tentatively predicted automation would impair conflict detection performance and that this impairment would be greater for poorer multi-taskers.

Methods

Participants

University undergraduate students (N = 125) were recruited from a psychology research participation pool. Participants provided informed consent and received AUD$10 and partial course credit. Data from nineteen participants were excluded: ten because they did not complete at least one task and nine because of corrupt data files. The final sample consisted of 106 participants (52 males, 54 females, M age = 21.23, range = 18 – 45). This research complied with the tenets of the Declaration of Helsinki and this project was approved by the Human Research Ethics Committee of The University of Western Australia.

Measures

Multi-tasking – Attentional Blink Task (AB*;* Chun & Potter, 1995): On each trial, participants were presented with a stream of 18 items, including distractors (digits from 2-9 and abstract symbols) and two target letters (excluding the letters I, L, O, Q, U, V, X). Each item was presented for 100ms. Targets (T1 and T2 respectively) were separated by zero (lag 2 = 200 ms),

two (lag 3 = 300 ms), or seven (lag 8 = 800 ms) distractors. At the end of each trial, participants identified the letters as accurately as possible using the keyboard, guessing if unsure. There was no time constraints for responding and there was an untimed break between blocks. Participants completed 120 trials (2 blocks × 60 trials divided equally across lags). The AB magnitude was calculated by subtracting the mean T2 given T1 accuracy at lags 2, 3 and 8. A smaller AB, or smaller difference in accuracy between lags indicated better working memory and filtering of distractors when encoding targets (relating to part 3 of the multi-tasking definition by Chérif et al., 2018). Thus a smaller AB reflected better multi-tasking. Task instructions to participants can be found in supplementary materials.

Multi-tasking – Psychological Refractory Period (PRP; Van Selst et al., 1999): On each trial, participants were presented with a randomly chosen symbol (from the set: #, &, @, or %) for 200-600ms, a short (200ms) or long (1000ms) blank interval, and a randomly chosen (from a set of four) complex tone for 200ms. Participants identified the symbol and tone as quickly and accurately as possible with responses for each of these two types of tasks mapped to keys on for the right hand (H, J, K or L) and left hand (A, S, D, or F). Participants completed 128 practice trials (16 per combination, mappings counterbalanced) to learn the keyboard mappings corresponding to the four possible symbols and four complex tones before they completed 160 experimental trials (5 blocks × 32 trials). Shorter RTs under each condition reflected faster response selection, task switching and dual task performance which are central to multi-tasking (parts 1 and 2 of multi-tasking definition of Chérif et al., 2018). Task instructions to participants can be found in supplementary materials.

Multi-tasking – Dual Response Selection task (Dual task; Dux et al., 2009): On each trial, participants completed either a single-task condition where they were presented with a single shape (hexagon or triangle) or complex tone (one of two abstract tones), or a dual task condition where they were presented with a one shape and one tone simultaneously. Participants identified the symbol and/or tone as quickly and accurately as possible using the keyboard letters assigned to left hand (A, S) and assigned to right hand (L, K) without making simultaneous keypresses (which would represent a lack of differentiation between stimuli – each hand responded to different stimuli thus represented a different task). Participants completed 24 practice trials to learn the keyboard response mappings (12 for each stimulus type; mappings counterbalanced) before they completed 144 experimental trials (4 blocks × 36 trials). This task used RT as a measure of response selection in the single stimuli and dual stimuli conditions, thus reflecting the executive processing cost of dual-tasking, which is central to multi-tasking (Chérif et al., 2018 – part 1 of multi-tasking definition). Lower RTs in the dual task condition thus reflected a lower cost of processing two stimuli concurrently. Task instructions to participants can be found in supplementary materials.

Performance – Air Traffic Control Task (Fothergill et al., 2009): Participants viewed an en-route sector display (Figure 2.1). Individual aircraft were represented by a circle attached to a data block, which provided the call sign, speed, aircraft type and altitude. A projection line indicated aircraft heading. Trials commenced with 7 or 8 aircraft presented at various locations on flight paths (indicated by the black lines) which followed designated flight paths before departing the sector. Aircraft positions updated every second and new aircraft entered the sector throughout the trial (average 1.5 per min). All aircraft travelled at a constant cruising altitude unless the participant intervened to resolve a conflict (see below).

Participants had 20 seconds to 'accept' aircraft approaching (press 'A' key, then click on aircraft) and 'hand-off' aircraft departing the sector (press 'H' key, then click on aircraft) the sector. In the automation (low–DOA) condition aircraft requiring acceptance or hand-off' (collectively called 'actioned' from here on) flashed and changed colour (blue for acceptances, orange for hand-offs) for 20s prior to entering or exiting the sector. Flashing stopped and colour returned to green once inside the sector or when actioned by the participant. This type of assistance was defined as low-DOA based on the general definition of such DOA from Wickens (2000) and has been used in several ATC studies previously (but not explicitly called low-DOA; see Loft et al., 2016; Tatasciore et al., 2020). In the manual condition aircraft did not flash or change colour prior to entering or exiting the sector. Participants' RT was recorded within this 20s timeframe prior to crossing sector boundary. Accuracy was calculated as a percentage of successful acceptances and hand-offs performed.

Additionally, participants had to detect and resolve conflicts between aircraft pairs that would violate five nautical mile lateral and 1000ft vertical separation in the future (click aircraft altitude triggering a prompt box requesting them to select the second conflict aircraft to initiate an altitude change for one of the aircraft). Aircraft pairs which were not intervened with and conflicted turned yellow when minimum separation was violated, then returned green once separation was re-established. Participants' RT for conflict detection was recorded from the time the second aircraft in a pair entered the sector (i.e., crossed the boundary or had been accepted) until the conflict had been actioned by the participant. If the conflict occurred (i.e., was not detected) the conflict time (i.e., time when the two aircraft first breached separation standards) was recorded. Only RTs for correctly intervened conflicts were included for data analysis. Accuracy was calculated as a percentage of the successful conflicts performed. Near misses were also recorded (i.e., successfully not intervening) but were not analysed. Participants completed two ATC conditions (low-DOA, manual) in counterbalanced order. Each condition contained: 44-47 aircraft to be accepted, 45-47 to be handed-offs (the numbers varied from person to person), 10 pairs of aircraft that would conflict if no intervention occurred, and six near-miss pairs of aircraft which passed closely but did not violate separation.

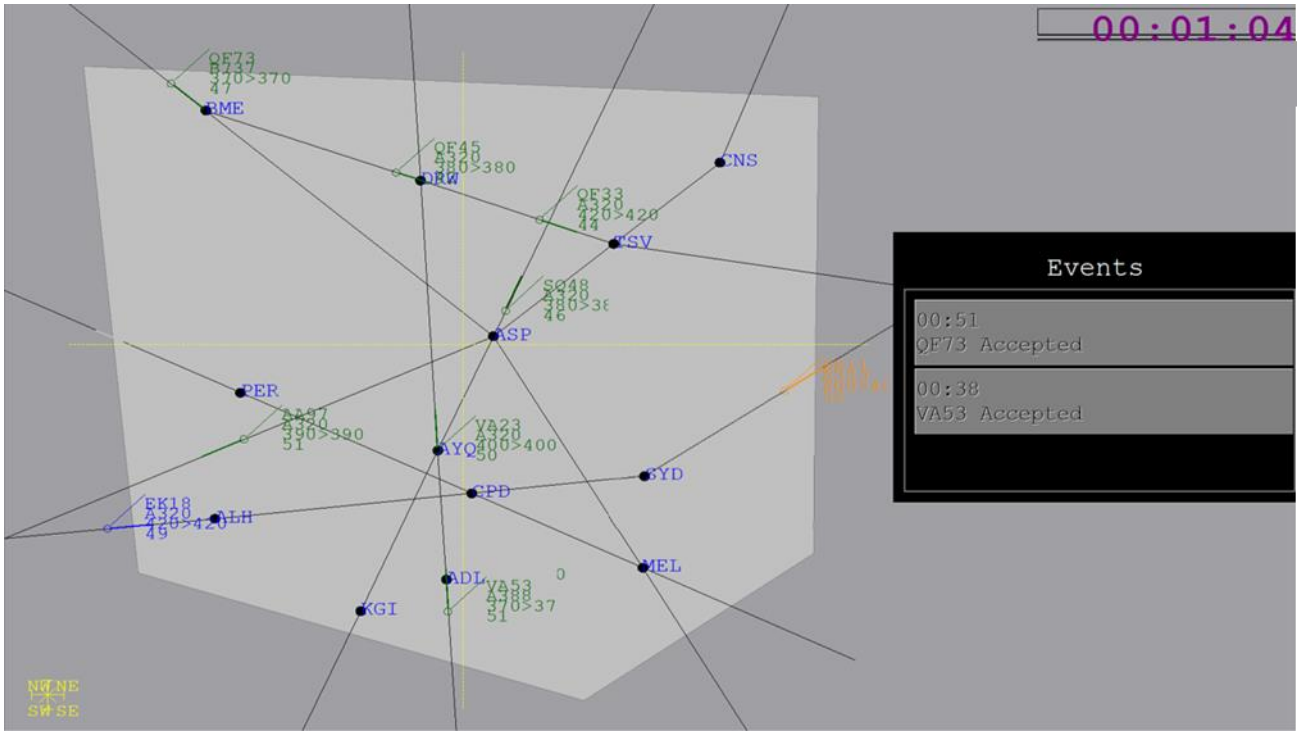


Figure 2.1: Air traffic control sector display. In the low-DOA condition (shown here), aircraft flashed blue when they approached the participants' sector (light grey polygon) indicating they required acceptance (i.e., EK18), and flashed orange indicating they required hand-off (i.e., EK11). In manual condition, all aircraft remained green. Actions performed by the participant were logged in the 'Events' box on the right of screen.

Situation Awareness: SA was measured using a modified Situation Present Awareness Method (SPAM; Durso et al., 2004). Every 2-3 minutes, participants received a visual 'Ready for Question?' prompt (18 queries per condition). They were instructed to click on this prompt within a ten-second window when their workload permitted. The ATC task was then paused but the display was not blanked. One SA query was then presented in a box at the top right of the display along with four response options (see Table 2.1 for examples). Participants were instructed to click a response option as quickly and accurately as possible.

Table 2.1. Examples of types of SA queries used in the ATC Task

Type of prompt	SA queries
Location	Are aircraft EK23 and aircraft AA31 currently located in the NW, NE, SW, or SE quadrant?
Flight level	What aircraft is on the same flight level as aircraft SQ57?
Conflict	In which quadrant will the next loss of separation take place within the next 30 secs if no action will be taken?
Acceptance	What is the flight level of the aircraft that you last accepted in to the sector in the NE quadrant?
Conflict	What aircraft is on the same flight level as aircraft AA37 in the SW quadrant?
Location	Will aircraft SQ79 and aircraft VA32 cross path in the NW, NE, SW, or SE quadrant?
Acceptance	How many aircraft needed accepting within the last 30 secs in the SW quadrant?
General	Which quadrant currently has the most traffic/number of aircraft?
Flight level	What is the flight level of the two aircraft that you last accepted in to the sector?
Location	Closest to which waypoint will aircraft AA83 and aircraft NZ55 cross path?
Location	What is the next waypoint that QF40 has to cross?
Speed	Are aircraft VA28 and SQ56 travelling at the same speed?
General	Which quadrant currently has the most traffic/number of aircraft?
Conflict	What common waypoint will aircraft SQ51 and aircraft QF81 both pass?
Flight level	What is the flight level of the aircraft that you last accepted on to the EK route?

Workload: Following the methodology of Strybel et al., (2016) and Loft et al., (2016), objective workload was operationalised as the time taken to respond to the "Ready for Question?" prompt that was presented prior to each SA query. In line with this, the time the operator takes to accept SA queries is often positively correlated with objective task load (Loft, Sadler, et al., 2015). Participants were not informed that they would be assessed on latency to answering SA queries so that a workload management strategy in which response to answering this prompt can be delayed in a manner that benefits workload.

Subjective workload was assessed with a prompt ('How would you rate your workload throughout the task?') presented once after each condition, which participants responded to by clicking on a graphical 7-point scale ranging from 1 (Very Low) to 7 (Very High). This item is based on the Air Traffic Workload Input Technique (ATWIT) which been used in previous ATC studies (Manning et al., 2001), however unlike usual application of the ATWIT which is typically presented periodically throughout a task, this measure was presented once at the end of an ATC scenario so as not to interrupt the task. This is a preliminary measure of workload (more extensive multi-dimensional measures of workload were used in Chapters 3 and 4 of this thesis).

Other Measures: Participants completed a demographics questionnaire (including age, gender, video gaming experience, previous ATC experience), and after the cognitive tasks a 9-item subscale from the Raven's Standard Progressive Matrices (RSPM; Bilker et al, 2012) to assess non-verbal intelligence. This was investigated as previous research has found general

intelligence is positively related to individual differences in multi-tasking, particularly relating to processing and storage components of working memory (Colom et al., 2010). After each condition, participants completed a nine-item measure of boredom and fatigue containing items from the Multidimensional State Boredom Scale (Fahlman et al., 2013) and the Fatigue scale (Chalder et al., 1993), the mental toughness scale (Gucciardi et al., 2015), a questionnaire requiring participants to rank their perceptions of the importance of the three ATC tasks (7-point Likert scale ranging from 'Not important' to 'Extremely important'). These were not central to the questions of this thesis and was not analysed as part of the thesis.

Procedure

All tasks were completed on a Windows PC, with a BENQ 24$^{\text{-inch}}$ monitor, providing a screen resolution of 1920 × 1080 running at 100Hz refresh rate. Participants first completed demographics, followed by a brief instructional overview of the cognitive tasks (written and oral; see supplementary materials for instructions provided for each task), and then the three cognitive tasks - the PRP task (20 minutes), the AB Task (10 minutes) and the Dual-Single task (10 minutes) - in counterbalanced order. Participants then completed a 20-minute training video in PowerPoint with voiceover which explained the aspects of the task, the difference between manual and automation conditions and the nature of the SA queries. No task priority was provided. The training video concluded with 10 multi-choice questions which provided feedback (showed correct answer) to ensure comprehension. The training video was followed by a 20-minute practice session of the manual condition, which provided participants with feedback in the form of the total number of successful acceptances, hand-offs and conflicts detected. Participants then completed two 30-minute experimental conditions in counterbalanced order (i.e., either manual then automation, or automation then manual conditions). After the testing session, participants were debriefed, given the opportunity to ask additional questions, and received remuneration.

Results

Data cleaning

For all cognitive measures, mean response time (RTs) were calculated using only trials with correct responses and RTs longer than 150ms and less than 3 SD above each participant's mean RT. This was the only data cleaning applied to the Dual task (Table 2.2). This resulted in exclusion of data from 1.2% of Dual task trials. Specific data cleaning procedures for the other cognitive tasks in addition to this are described below.

PRP Task: In addition to the data cleaning described above, PRP calculations omitted RTs if there was less than 50ms between keyboard responses (Tombu & Jolicœur, 2005) as these 'grouped responses' do not reflect sequential decision making (Ulrich & Miller, 2008) assumed by the response bottleneck and thus do not reflect 'multi-tasking' by responding to two stimuli (Lien

et al., 2003). Mean RTs were calculated separately for each stimulus type (visual or auditory: PRP; Table 2.3). Data cleaning resulted in the exclusion of data from 14.8% of PRP trials.

AB task: Accuracy for T2 was calculated only for trials in which the first target (T1) was correctly identified. Mean T1 and T2 accuracy were calculated separately as a function of lag (Raymond et al., 1992; see Table 2.4). No data was excluded from the AB.

ATC task: Mean RTs were calculated using only trials with correct responses and RTs longer than 150ms and less than 3 SD above each participant's mean RT (see Table 2.5). This resulted in the exclusion of 2.66% of acceptances, 3.11% of hand-offs, and 0% of conflict trials. Adjusted RTs were then calculated separately for acceptance, hand-off, and conflict detection by dividing mean RTs by mean accuracy, providing a single composite measure which combines accuracy and RT to account for speed-accuracy trade-offs (Liesefeld & Jancyzk, 2019; Visser et al., 2015).

SA: Mean RTs were calculated using only SPAM queries with correct responses and RTs longer than 150ms or less than 3 SD above each participant's mean RT (see Table 2.5). This resulted in the exclusion of 1.43% of SA queries. Additionally, responses from one SA query were omitted due to technical error. Adjusted RTs were calculated by dividing RTs by mean accuracy.

Objective Workload: Mean RTs were calculated as the time taken to respond to the "Ready for Question?" prompt prior to each SPAM question, but only on trials in which the subsequent SA query was answered correctly.

Table 2.2. Mean percentage of targets correctly identified and mean response time (in milliseconds) in the Dual Response Selection Task, separated by target type, with standard deviation in parentheses.

Trial Type	Shape		Sound	
	Accuracy	RT	Accuracy	RT
Single task	93.63 (7.42)	753 (158)	92.35 (8.45)	848 (135)
Dual task	91.29 (8.70)	927 (188)	91.29 (8.70)	1086 (170)

Table 2.3. Mean percentage of targets correctly identified and mean response time (in milliseconds) in the PRP task, separated by target type, with standard deviation in parentheses.

Inter-target interval	Sound trials		Visual trials	
	Accuracy	RT	Accuracy	RT
200ms	98.30 (1.32)	2080 (687)	98.13 (1.27)	1625 (630)
1000ms	97.46 (2.21)	1550 (567)	96.25 (10.19)	1691 (728)

Table 2.4. Mean percentage of targets correctly identified in the AB task with standard deviation in parentheses.

	T1	T2|T1
Lag 1	84.60 (10.17)	87.39 (9.97)
Lag 3	88.30 (10.49)	42.75 (20.33)
Lag 8	88.02 (9.16)	82.53 (12.13)

Table 2.5. Mean performance in the ATC task (acceptance, hand-off, conflict detection) and SPAM task as a function of condition. Accuracy is the percentage of correct trials. RTs are response times on correct trials. Standard deviations are in parentheses.

	Manual		Low-DOA	
Task	Accuracy	RT	Accuracy	RT
Acceptance	94.65 (10.02)	4.20 (1.43)	99.32 (4.98)	2.35 (0.86)
Hand-off	94.25 (10.85)	3.75 (1.43)	99.53 (3.69)	2.58 (1.01)
Conflict detection	96.60 (10.50)	97.02 (27.67)	95.47 (32.09)	93.27 (32.09)
SPAM	87.59 (12.58)	16.95 (5.38)	87.73 (10.84)	17.14 (6.80)

Latent factor analysis

Exploration of the multi-tasking factor began with the intention of including PRP, Dual cost and AB performance metrics together in a factor. One measure from each task was intended to ensure the latent factor was equally representative of the aspects of multi-tasking being assessed. A preliminary Exploratory Factor Analysis (EFA) conducted with combinations of three variables with (at least one from each cognitive task) resulted in loadings for all three tasks of less than .40. The AB did not load with the other tasks (highest loading = .08) and was excluded from the final factor.

The loading between the PRP (Short PRP RT) and Dual (Dual RT) tasks were large and significant and theoretically represented multi-tasking aspects including task switching, response selection, encoding and retrieval. These variables corresponded to performance at the shortest inter-target interval in the PRP task and the simultaneous stimulus presentation in the Dual task which reflects the more difficult multi-tasking requirements in these tasks. These were included in the Exploratory Factor Analysis subsequently conducted. The EFA was conducted with Principal Axis Factoring extraction method to maximize loadings, and Direct Oblimin as measures were

highly correlated. The Kaiser-Meyer-Olkin (KMO = .50) and Bartlett's Test of Sphericity ($p<.05$) showed the data was suitable for the analysis. The result of the Scree Plot and the Kaiser's criteria (eigenvalue > 1) confirmed one factor could be extracted from the tasks. The extracted factor formed from these two variables explained 52.9% of the variance (both loading =.73; uniqueness = .50).

A multi-tasking factor score for was created for each participant using their Bartlett scores from the EFA (multiplied by -1 so that higher multi-tasking factor scores represented better multi-tasking ability; Figure 2.2; Bartholomew et al., 2009). These factor scores isolate shared variance on a factor across constituent tasks, producing an unbiased estimate using maximum likelihood estimates (DiStefano et al., 2009) and were used as the predictor for subsequent analyses.

Linear mixed models

To examine whether the effects of automation on outcome measures differed as a function of multi-tasking ability, a series of linear mixed effects models (LMM) were conducted using the lme4 plugin (Bates et al., 2014) for R (RCore Team, 2015). Data were entered in a nested form to account for the within-subject design. Condition was entered as a fixed factor with two levels – low-DOA or manual – and multi-tasking score was entered as a fixed continuous effect. To examine the interaction between multi-tasking and condition, an interaction term was entered as a fixed effect. Participant number was entered as a random effect (intercept) along with condition as a random effect (slope) to control for within-subject variability differing across condition. The model predictors were chosen *a priori* to test the hypotheses of interest.

An LMM was fitted for each dependent variable. Model significance was assessed using a Chi-Square test to compare to a null model which included only the random effects and the intercept (see supplementary materials Table 2). Unless otherwise stated, models significantly differed from the null. Assumptions of linearity, absence of autocorrelation, absence of influential cases and multicollinearity were satisfied (see supplementary materials Assumption plots). Homoscedasticity was at an acceptable level (i.e., evenly distributed) for all models as there was only two instances of participants with higher level variability. The assumption of normality of residuals was violated in all models, however LMM have been found to be robust to violations of this assumption (Knief & Forstmeier, 2021; see review by Schielzeth et al., 2020). Fixed factor effects and model fit statistics are presented in Table 2.6.

Models were initially conducted with RSPM, previous experience with ATC, and counterbalance order as covariates to control for potential effects of non-verbal intelligence, task order, or practice effects. The only significant covariate was the RSPM, which predicted conflict detection such that higher non-verbal intelligence related to greater performance (see

supplementary materials Table 3). Inclusions of these covariates did not influence any of the results reported below. Thus, for brevity, these covariates are not discussed further.

Standardised betas are presented in Table 2.6 to allow comparison of effect sizes across studies and measures. Negative betas indicate a negative effect of multi-tasking, meaning that adjusted RT decreased as multi-tasking scores increased. Likewise negative betas for condition indicate adjusted RTs decreased in the automation compared to the manual condition. Decreased adjusted RTs reflect better performance (i.e., more efficient).

ATC Performance: For acceptances and hand-offs (see Figure 2.2 for mean performance and Table 2.6 for effect betas), the significant effect of multi-tasking indicated that participants with higher multi-tasking scores performed better, while the significant effect of condition indicated that performance was better in the low-DOA compared to the manual condition. The significant condition by multi-tasking interaction indicated that improvements in the low-DOA condition increased as multi-tasking scores decreased. For conflict detection, the significant effect of multi-tasking indicated that participants with higher multi-tasking scores performed better. There was no significant main effect of condition or interaction.

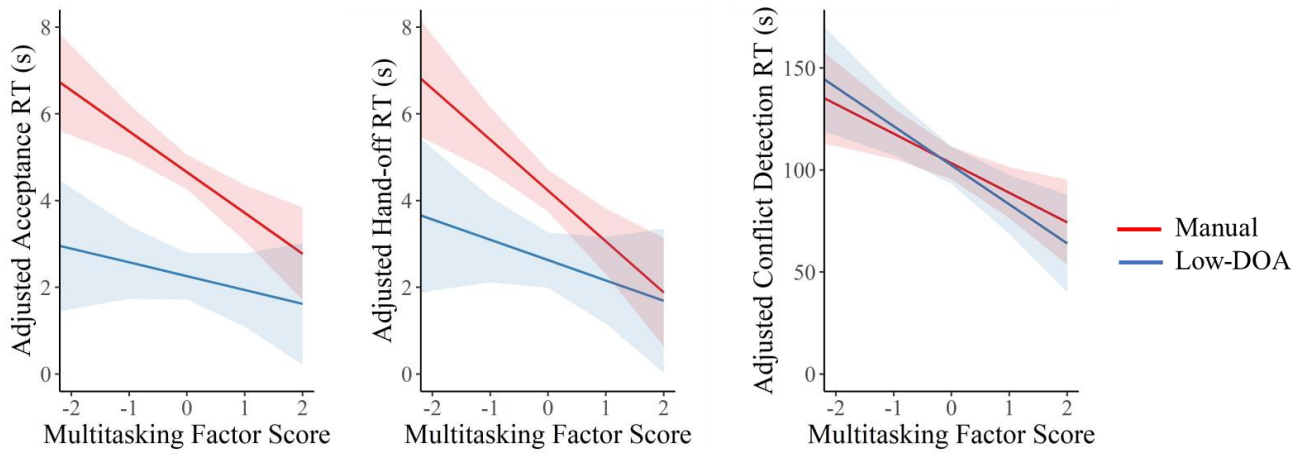


Figure 2.2. Aircraft acceptance (left), hand-off (centre) and conflict detection (right) LMM slopes and intercepts. Performance measured via adjusted response time (seconds) against multi-tasking score (standardised), where a lower adjusted RT reflects better performance and higher multi-tasking scores reflect better multi-tasking ability. Shaded area represents 95% CI.

Situation Awareness: The significant effect of multi-tasking (see Figure 2.3 for mean performance and Table 2.6 for effect betas) indicated that participants with higher multi-tasking scores had better SA. There was no significant main effect of condition or interaction.

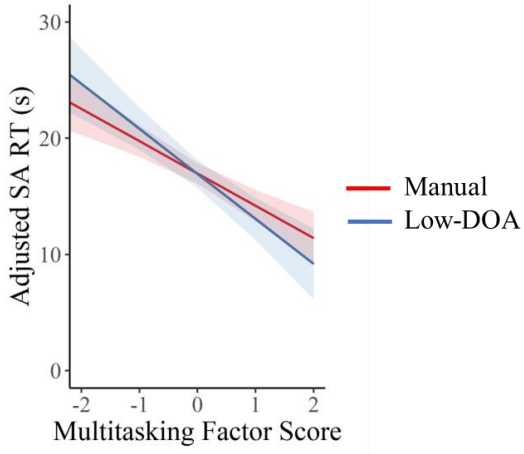


Figure 2.3. Situation Awareness LMM slope and intercept. Performance measured in adjusted response time (seconds) against multi-tasking score (normed). Black line represents manual condition and grey line represents low-DOA condition. Shaded area represents 95% CI.

Objective Workload: A significant effect of multi-tasking (see Figure 2.4 for mean performance and Table 2.6 for effect betas) indicated that participants with higher multi-tasking scores were quicker to accept SA prompts, suggesting lower workload. There was no significant main effect of condition (manual: mean = 2.19, SD = 1.06; automation: mean = 2.09, SD = 0.85) or interaction.

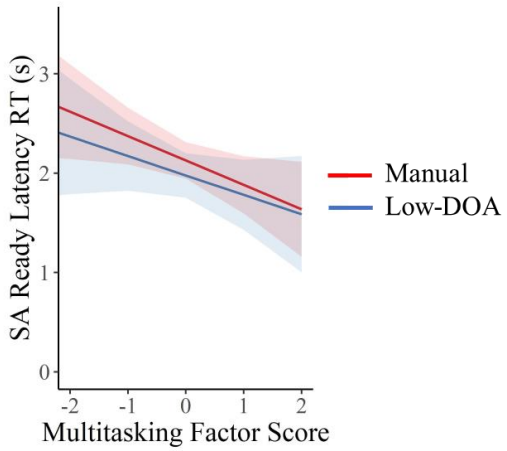


Figure 2.4. Objective workload measured by response latency to SPAM 'ready' prompt (measured in seconds) LMM slope and intercept. Black line represents manual condition and grey line represents low-DOA condition. Shaded area represents 95% CI.

Table 2.6. Linear mixed effects standardised coefficient estimates, standard deviations in parantheses and model fit summaries.

Parameter	Acceptances Adjusted RT	Hand-offs Adjusted RT	Conflict Detection Adjusted RT	SA Question Adjusted RT	Objective Workload
Condition	-0.34 ***	-0.24***	-0.01	$-3.42x10^{-3}$	-0.04
	(0.19)	(0.22)	(3.34)	(0.49)	(0.08)
Multi-tasking	-0.17 ***	-0.20***	-0.21 **	-0.28 ***	-0.10 *
	(0.25)	(0.29)	(4.86)	(0.52)	(0.11)
Condition × Multi-tasking	0.08 **	0.09**	-0.02	-0.05	0.01
	(0.22)	(0.26)	(4.03)	(0.59)	(0.10)
Observations	9069	8997	2036	3264	3204
Log Likelihood	-21954.91	-22602.09	-11200.81	-11761.33	-6155.25
AIC	43923.8	45218.18	22415.62	23536.66	12324.51
BIC	43973.6	45267.92	22454.95	23579.29	12367.02
R2 (conditional)	0.48	0.47	0.35	0.25	0.25
R2 (marginal)	0.12	0.07	0.04	0.08	0.01

Note: *** $p < .001$, ** $p < .01$, * $p < .05$. The estimates represent standardised coefficient estimates for fixed effects in the linear mixed model on the trial level specifying a random intercept and slope per participant. P-values were computed using the Wald approximation. R2 conditional is the model's total explanatory power. R2 marginal is the expanatory power of the fixed effects alone. Observations are the number of data points (trials) analyzed. Log likelihood is a measure of fit describing how well a parameter explains the observed data. AIC is Akaike Information Criterion, an extimator of prediction error. BIC is Bayesian Information Criterion, a measure for model comparison or selection.

Subjective Workload: Data was unsuitable for LMM analysis as with exactly two data-points per participant it is not possible to model the random slope for each participant. Instead, participants were divided into three equal groups based on multi-tasking ability factor scores (bottom 33%, middle 33%, top 33%). A one-way ANOVA was conducted to validate these groups, showing there was a significant difference between multi-tasking group means. Workload ratings were then analysed using a 3 (multi-tasking ability: low, medium, high) × 2 (condition: automation, manual) analysis of variance. As shown in Figure 2.5, this yielded only a significant effect of condition, $F(1,103) = 52.71$, $p < .001$, $\eta^2 = 0.10$, with lower subjective workload in the automation condition. There was no significant effect of multi-tasking ability or interaction ($F < 2.30$, $p > .09$, $\eta^2 < .03$).

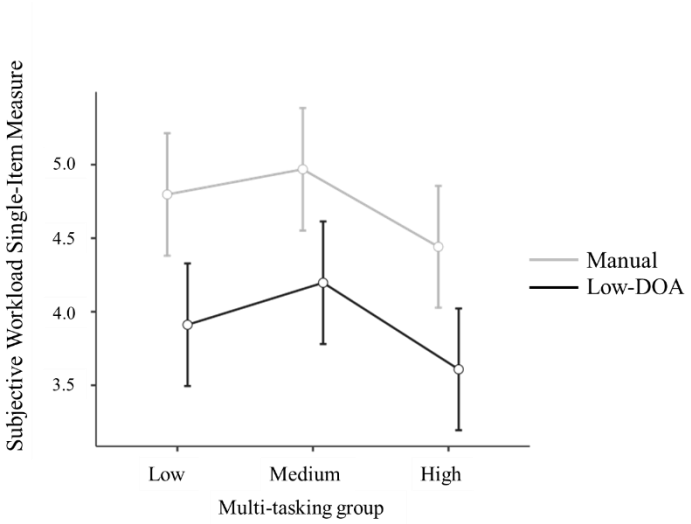


Figure 2.5. Subjective workload measure across three levels of multi-tasking ability and two conditions of the ATC. Error bars represent 95% confidence intervals around the estimate of marginal means (between subjects).

Discussion

Previous studies have focused on understanding how varying DOA modulates the impact of automation on outcome measures. Here, this chapter asked a complimentary but distinct question: 'do individual differences in multi-tasking ability modulate the impact of automation on outcome measures?' To this end, this chapter examined whether participant's multi-tasking ability modulated the benefits of using low-DOA in a simulated ATC task. Across outcome measures of ATC performance, workload and SA, this chapter found automation improved performance on the acceptance and hand-off tasks and led to lower subjective workload, but the low-DOA applied to these ATC tasks did not benefit SA or objective workload and had no impact on the non-automated conflict detection task. This chapter also found effects of multi-tasking ability on all measures except subjective workload, and support for the predicted interaction between automation and multi-tasking on the acceptance and hand-off tasks. However, this interaction between automation and multi-tasking did not extend to workload, SA, or the non-automated conflict detection task.

Individual differences can modulate the benefits of automation

The predictions about the benefits of multi-tasking were two-fold: first, greater multi-tasking ability should be associated with superior performance, SA, and workload outcomes. Second, multi-tasking ability should interact with automation, such that automation benefits should be greater for poorer multi-taskers. The first prediction was strongly supported as most outcome measures improved with increased multi-tasking ability, with the sole exception of subjective workload. In sum, the effect of multi-tasking ability on performance and SA is consistent with the well-established findings that cognitive overload leads to reduced task

performance and higher workload (Saqer & Parasuraman, 2014) as well as established links between multi-tasking and task performance (Wickens et al., 2015). Moreover, it expands on this work by showing normal variations in cognitive ability can affect performance and SA in the context of dynamic and complex tasks.

The second key prediction that multi-tasking would modulate the benefit of automation on outcome measures was partially supported. There were larger benefits of automation for poorer multi-taskers than better multi-taskers on performance in both the acceptance and hand-off tasks. This novel finding suggests that low-DOA freed-up more cognitive resources for participants with poorer multi-tasking ability, and those additional resources resulted in proportionally larger benefits to task performance compared to manual task performance. This outcome is broadly consistent with earlier studies suggesting individual differences in spatial ability, information processing and working memory capacity (Jipp & Ackerman, 2016; Wright et al., 2018) can modulate the impact of increasing DOA. However, a key difference is that whereas earlier work looked at how individual differences modulated the impact of *different* DOAs *within the same individual*, this chapter study shows that individual differences modulate the impact of the *same* DOA *across different individuals*. This distinction highlights the importance of considering individual differences when deploying automation across a group of operators, as the present findings indicate its effectiveness could vary with differences in cognitive ability.

That said, it is notable that multi-tasking ability did not interact with automation for other outcome measures. In the case of SA, this is likely linked to the failure to detect a benefit of automation on SA. As noted earlier, past studies that have shown inconsistent effects of automation on SA (e.g., Kaber & Endsley, 2004). Moreover, a closer examination of the literature indicates no studies using SPAM methodology have shown automation-linked SA benefit (e.g., S. Chen et al., 2017). This may reflect the fact that SPAM assesses SA in terms of an operator's ability to find task-relevant information in the environment rather than their ability to retrieve information from a detailed mental model as is the case with other SA probe techniques (Endsley, 2019a; 2019b). In turn, information retrieval from a display (rather than from memory) is likely to be less cognitively demanding and may not benefit as much from the type or quantity of cognitive resources freed-up by low-DOA.

With respect to workload, this chapter found diverging patterns of results, as automation only benefited subjective workload, and multi-tasking ability only benefited objective workload. This suggests that the failure to detect an interaction between multi-tasking ability and automation on either measure of workload may be the result of a relatively weak relationship between the availability of cognitive resources and workload. This could have arisen because workload was relatively moderate (subjective ratings ranging from 3.90-4.73 on a 7-point scale) and thus not sufficiently taxing on cognitive resources to see consistent benefits of either automation or multi-tasking. Of course, this explanation is speculative, and it should be noted that the measures in this

chapter were preliminary examinations of workload. Subsequent chapters re-examined workload with additional measures of subjective workload to validate and expand the scope of these results.

Effects of multi-tasking and low-DOA on non-automated task performance

An additional issue probed here was the potential impact of multi-tasking and low-DOA on the non-automated conflict detection task. As expected, the results indicated that superior multi-tasking ability was linked to better conflict detection. However, low-DOA had no impact on conflict detection relative to manual performance and did not interact with multi-tasking ability. This result was unexpected as this chapter predicted that automation of the acceptance task would potentially impair conflict detection due their dependence on common display information. One possible explanation for the null findings is that the degree of information overlap between the tasks was less than anticipated. This seems plausible because although automation aided initial detection of aircraft entering the sector, it did not aid subsequent aircraft tracking which was also important for conflict detection. With less dependence on common display information, one would also expect the predicted cost of automation to decrease, thus accounting for the non-significant impact of automation on conflict detection seen here.

Limitations

Several limitations of the present chapter should be highlighted. First, the multi-tasking latent factor was formed from only two cognitive tasks. This outcome is not uncommon in published studies (e.g., Bradshaw et al., 2009), but may be less robust as it can yield under-defined models (Draheim et al., 2019). Thus, it would be desirable in future work to administer additional multi-tasking assessments to ensure a latent factor can be constructed from at least three tasks to verify the robustness of the present results.

Second, these findings were obtained using novice participants rather than experienced, highly trained operators. Research suggests that experienced operators interact with automation systems differently than novices (Jamieson & Skraaning, 2018), which could modulate the impact of automation. Further, experienced operators have been extensively trained and have typically self-selected into roles requiring high-level cognitive skills. This could constrain the magnitude and influence of individual differences in cognitive ability amongst such experienced operators. For these reasons, it would be desirable to replicate the present findings with experienced operators to examine their generalisability and boundary conditions.

Finally, task fidelity has also been noted as a moderating factor in automation outcomes (Jamieson & Skraaning, 2020). The present ATC task is representative of several features of real ATC and thus reflects some of the multi-tasking requirements faced by expert operators. However, the task can be considered medium fidelity as real ATC requires an even greater number of tasks with higher multi-tasking requirements than is simulated here.

Practical implications and future directions

One potential practical outcome of these findings is that low-DOA, which is often less beneficial to performance than high-DOA (Onnasch et al., 2014), can yield considerably greater benefits for individuals with poorer cognitive abilities (such as multi-tasking). A valuable question for future research is to determine the range of cognitive abilities to which the present findings may generalise. Another important question is whether individual differences in cognitive abilities would also modulate the impact of high-DOA. This is a particularly interesting issue because high-DOA typically benefits performance and workload significantly, but also reduces SA and can lead to "out-of-the-loop" problems when automation fails, and operators are required to assume manual control (Kaber, 2018). Thus, it is possible that while higher-DOA could be even more beneficial for individuals with poorer cognitive abilities, costs to SA and the consequences of automation failure could also be greater.

Another practical implication of this chapter concerns the continuing debate about the utility of assessing specific cognitive predictors beyond widely used measures of general cognitive ability (Morgan et al., 2013). As noted in the results, preliminary analyses found non-verbal intelligence did not modulate the impact of multi-tasking across outcomes. Moreover, the correlation between non-verbal intelligence and multi-tasking ability was relatively modest (r = .34). These findings suggest that the potential contribution of multi-tasking to outcomes was distinct from that attributable to its overlap with intelligence, and bolsters recent calls (Barron & Rose, 2017) for greater use of specific cognitive abilities measures to augment general ability measures when predicting future outcomes in areas such as training and on-the-job performance.

Conclusion

The limitations described above notwithstanding, determining what automation support to provide to operators in future design of workplace systems could consider individual differences. This chapter's results showed participants with poorer multi-tasking ability generally had poorer outcomes but demonstrated significant greater performance benefits from low-DOA than individuals with higher multi-tasking ability. This suggests that current design approaches that provide everyone with the same automation could consider the potential benefits of a "many-sizes-fits-all" approach that provides automation flexibly depending on capabilities and need.

References

Bartholomew, D. J., Deary, I. J., & Lawn, M. (2009). The origin of factor scores: Spearman, Thomson and Bartlett. *British Journal of Mathematical and Statistical Psychology*, 62(3), 569-582.

Barron, L. G., & Rose, M. R. (2017). Multitasking as a predictor of pilot performance: Validity beyond serial single-task assessments. *Military Psychology*, 29(4), 316-326.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv* preprint arXiv:1406.5823.

Bender, A., Loft, S., Lipp, & Visser, T.A.W. (2018). *Advancing our understanding of warfighter cognition: Development of a "cognitive profiling" tool to enhance situation awareness.* Defence Science and Technology (DST) Group Human Performance Research Network (HPRnet).

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment*, *19*(3), 354-369.

Bradshaw, K., Roesch, S. C., Nadler, K., Ehrhart, M. G., Chung-Herrera, B. G., & Ehrhart, K. H. (2009). Testing the latent factor structure and construct validity of the Ten-Item Personality Inventory. *Personality and Individual Differences*, 47(8), 900–905.

Bühner, M., König, C. J., Pick, M., & Krumm, S. (2006). Working Memory Dimensions as Differential Predictors of the Speed and Error Aspect of Multitasking Performance. *Human Performance*, *19*(3), 253–275.

Cak, S., Say, B., & Misirlisoy, M. (2020). Effects of working memory, attention, and expertise on pilots' situation awareness. *Cognition, Technology & Work*, 22(1), 85-94.

Calhoun, G., Draper, M., & Ruff, H. A. (2009). Effect of level of automation on unmanned aerial vehicle routing task. *Proceedings of the Human Factors Society annual meeting*, 197–201.

Chalder, T., Berelowitz, G., Pawlikowska, T., Watts, L., Wessely, S., Wright, D., & Wallace, E. P. (1993). Development of a fatigue scale. *Journal of psychosomatic research*, *37*(2), 147-153.

Chen, J. Y. C., & Terrence, P. I. (2009). Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multi-tasking environment. *Ergonomics*, *52*(8), 907–920.

Chen, S. I., Visser, T. A., Huf, S., & Loft, S. (2017). Optimizing the balance between task automation and human manual control in simulated submarine track management. *Journal of Experimental Psychology: Applied*, *23*(3), 240.

Chérif, L., Wood, V., Marois, A., Labonté, K., & Vachon, F. (2018). Multi-tasking in the military: Cognitive consequences and potential solutions. *Applied Cognitive Psychology*, *32*(4), 429–4.

Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 109–127.

Colom, R., Martínez-Molina, A., Shih, P. C., & Santacreu, J. (2010). Intelligence, working memory, and multitasking performance. *Intelligence*, 38(6), 543-551.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation*, *14*(1), 20.

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, 145(5), 508.

Durso, F. T., Dattel, A. R., Banbury, S., & Tremblay, S. (2004). SPAM: The real-time assessment of SA. *A cognitive approach to situation awareness: Theory and application*, *1*, 137-154.

Duncan, J., Martens, S., & Ward, R. (1997). Restricted attentional capacity within but not between sensory modalities. *Nature*, 387(6635), 808-810.

Dux, P. E., Tombu, M. N., Harrison, S., Rogers, B. P., Tong, F., & Marois, R. (2009). Training Improves Multitasking Performance by Increasing the Speed of Information Processing in Human Prefrontal Cortex. *Neuron*, 63(1), 127-138.

Endsley, M. R. (1988a). Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 national aerospace and electronics conference* (pp. 789-795). IEEE.

Endsley, M. R. (1988b). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society annual meeting*, 32(2), 97-101. Sage CA: Los Angeles, CA: Sage Publications.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors*, *37*(1), 65-84.

Endsley, M. R. (2019a). A Systematic Review and Meta-Analysis of Direct Objective Measures of Situation Awareness: A Comparison of SAGAT and SPAM. *Human Factors*.

Endsley, M. R. (2019b). The Divergence of Objective and Subjective Situation Awareness: A Meta-Analysis, *XX*(X).

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, 42(3), 462-492.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, 37(2), 381-394.

Fahlman, S. A., Mercer-Lynn, K. B., Flora, D. B., & Eastwood, J. D. (2013). Development and validation of the multidimensional state boredom scale. *Assessment*, *20*(1), 68-85.

Fothergill, S., Loft, S., & Neal, A. (2009). ATC-labAdvanced: An air traffic control simulator with realism and control. *Behaviour Research Methods*, *41*(1), 118–127.

Gutzwiller, R. S., Wickens, C. D., & Clegg, B. A. (2019). The role of reward and effort over time in task switching. *Theoretical issues in ergonomics science*, *20*(2), 196-214.

Gucciardi, D. F., Hanton, S., Gordon, S., Mallett, C. J., & Temby, P. (2015). The concept of mental toughness: Tests of dimensionality, nomological network, and traitness. *Journal of personality*, *83*(1), 26-44.

Jamieson, G. A., & Skraaning, G. (2018). Levels of automation in human factors models for automation design: Why we might consider throwing the baby out with the bathwater. *Journal of Cognitive Engineering and Decision Making*, 12(1), 42-49.

Jamieson, G. A., & Skraaning, G. (2020). The absence of degree of automation trade-offs in complex work settings. *Human Factors,* 62(4), 516-529.

Jipp, M., & Ackerman, P. L. (2016). The Impact of Higher Levels of Automation on Performance and Situation Awareness. *Journal of Cognitive Engineering and Decision Making, 10*(2), 138–166.

Johnston, J. C., & Pashler, H. (1998). Attentional limitations in dual-task performance. Attention, 155-189.

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*(2), 113-153.

Kaber, D. B. (2018). Issues in Human–Automation Interaction Modeling: Presumptive Aspects of Frameworks of Types and Levels of Automation. *Journal of Cognitive Engineering and Decision Making, 12*(1), 7–24.

Kaber, D. B., Onal, E., & Endsley, M. R. (2000). Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. *Human Factors and Ergonomics in Manufacturing*, 10, 409-430.

Klapp, S. T., Maslovat, D., & Jagacinski, R. J. (2019). The bottleneck of the psychological refractory period effect involves timing of response initiation rather than response selection. *Psychonomic bulletin & review*, 26(1), 29-47.

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, 53(6), 2576-2590.

Körber, M., Weißgerber, T., Kalb, L., Blaschke, C., & Farid, M. (2015). *Prediction of take-over time in highly automated driving by two psychometric tests*, *82*(193), 195–201.

Lien, M. C., Schweickert, R., & Proctor, R. W. (2003). Task switching and response correspondence in the psychological refractory period paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, 29(3), 692.

Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs (?). *Behavior Research Methods*, 51(1), 40-60.

Loft, S., Bowden, V., Braithwaite, J., Morrell, D. B., Huf, S., & Durso, F. T. (2015). Situation awareness measures for simulated submarine track management. *Human Factors*, 57(2), 298-310.

Loft, S., Morrell, D., Ponton, K., Braithwaite, J, Bowden, V., & Huf, S. (2016). The impact of uncertain contact location on situation awareness and performance in simulated submarine track management. *Human Factors,* 58, 1052-1068.

Loft, S., Sadler, A., Braithwaite, J., & Huf, S. (2015). The chronic detrimental impact of interruptions in a simulated submarine track management task. *Human Factors*, 57(8), 1417-1426.

Manning, C., Mills, S., Fox, C., Pfleiderer, E., & Mogilka, H. (2001). The relationship between air traffic control communication events and measure of controller taskload and workload. *In the 4th USA/Europe Air Traffic Management R&D Seminar*. Santa Fe, NM.

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87.

Mansikka, H., Virtanen, K., & Harris, D. (2018). Dissociation between mental workload, performance, and task awareness in pilots of high performance aircraft. *IEEE Transactions on Human-Machine Systems*, 49(1), 1-9.

Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in cognitive sciences*, 9(6), 296-305.

Morgan, B., D'Mello, S., Abbott, R., Radvansky, G., Haass, M., & Tamplin, A. (2013). Individual differences in multitasking ability and adaptability. *Human Factors*, *55*(4), 776-788.

Nguyen, T., Lim, C. P., Nguyen, N. D., Gordon-Brown, L., & Nahavandi, S. (2019). A review of situation awareness assessment approaches in aviation environments. *IEEE Systems Journal*, *13*(3), 3590-3603.

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors*, 56, 476–488.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, *30*(3), 286-297.

Parasuraman, R., Cosenzo, K., & de Visser, E. (2009). Adaptive Automation for Human Supervision of Multiple Uninhabited Vehicles: Effects on Change Detection, Situation Awareness, and Mental Workload. *Military Psychology*, 21(2), 270.

Pashler, H. (1984). Processing stages in overlapping tasks: evidence for a central bottleneck. *Journal of Experimental Psychology: Human perception and performance*, 10(3), 358.

Pashler, H. (1992). Attentional limitations in doing two tasks at the same time. *Current Directions in Psychological Science*, 1(2), 44-48.

Pierce, R. S., Vu, K. P. L., Nguyen, J., & Strybel, T. Z. (2008, September). The relationship between SPAM, workload, and task performance on a simulated ATC task. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 52(1), 34-38. Sage CA: Los Angeles, CA: Sage Publications.

R Development Core Team (2015) R: *A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink?. *Journal of experimental psychology: Human perception and performance*, 18(3), 849.

Redick, B. T. S., Shipstead, Z., Meier, M. E., Montroy, J. J., Hicks, K. L., Unsworth, N., … Hambrick, D. Z. (2016). Cognitive predictors of a common multi-tasking ability: Contributions from working memory, attention control, and fluid intelligence. *Journal of experimental psychology: General* 145(11), 1473.

Redick, T. S. (2016). On the relation of working memory and multitasking: Memory span and synthetic work performance. *Journal of Applied Research in Memory and Cognition*, 5(4), 401-409.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, 49(1), 76-87.

Saqer, H., & Parasuraman, R. (2014). Individual performance markers and working memory predict supervisory control proficiency and effective use of adaptive automation. *International Journal of Human Factors and Ergonomics 55, 3*(1), 15-31.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., ... & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in ecology and evolution*, 11(9), 1141-1152.

Sethumadhavan, A. (2009). Effects of automation types on air traffic controller situation awareness and performance. *Proceedings of the human factors and ergonomics society 53rd annual meeting—2009 1 effects*, 1329–1333.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.

Strybel, T. Z., Vu, K. P. L., Chiappe, D. L., Morgan, C. A., Morales, G., & Battiste, V. (2016). Effects of NextGen concepts of operation for separation assurance and interval management on Air Traffic Controller situation awareness, workload, and performance. *The International Journal of Aviation Psychology*, *26*(1-2), 1-14.

Tatasciore, M., Bowden, V. K., Visser, T. A., Michailovs, S. I., & Loft, S. (2020). The benefits and costs of low and high degree of automation. *Human Factors*, *62*(6), 874-896.

Tatasciore, M., Bowden, V. K., Visser, T. A., & Loft, S. (2021). Should We Just Let the Machines Do It? The Benefit and Cost of Action Recommendation and Action Implementation Automation. *Human Factors*, 0018720821989148.

Taylor, R. M., & Selcon, S. J. (1990, October). Cognitive quality and situational awareness with advanced aircraft attitude displays. *In Proceedings of the Human Factors Society Annual Meeting*, 34(1), 26-30. Sage CA: Los Angeles, CA: SAGE Publications.

Thomas, M. L., & Russo, M. B. (2007). Neurocognitive monitors: toward the prevention of cognitive performance decrements and catastrophic failures in the operational environment. *Aviation, space, and environmental medicine*, *78*(5), B144-B152.

Tombu, M., & Jolicœur, P. (2005). Testing the predictions of the central capacity sharing model. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(4), 790.

Ulrich, R., & Miller, J. (2008). Response grouping in the psychological refractory period (PRP) paradigm: Models and contamination effects. Cognitive Psychology, 57(2), 75-121.

Van Selst, M., Ruthruff, E., & Johnston, J. C. (1999). Can practice eliminate the psychological refractory period effect?. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(5), 1268.

Vidullch, M. A., Ward, G. F., & Schueren, J. (1991). Using the subjective workload dominance (SWORD) technique for projective workload assessment. *Human Factors*, 33(6), 677-691.

Visser, T. A. W., Ohan, J. L., & Enns, J. T. (2015). Temporal cues derived from statistical patterns can overcome resource limitations in the attentional blink. *Attention, Perception, and Psychophysics*, *77*(5), 1585–1595.

Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and Automation Bias in the Use of Imperfect Automation. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *57*(5), 728–739.

Wickens, C. D., Gutzwiller, R. S., & Santamaria, A. (2015). Discrete task switching in overload: A meta-analyses and a model. *International Journal of Human Computer Studies*, *79*, 79–84.

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. B. (2010, September). Stages and levels of automation: An integrated meta-analysis. In *Proceedings of the human factors and ergonomics society annual meeting*, 54(4), 389-393. Sage CA: Los Angeles, CA: Sage Publications.

Wright, J. L., Chen, J. Y. C., & Barnes, M. J. (2018). Human-Automation Interaction for Multiple Robot Control: The Effect of Varying Automation Assistance and Individual Differences on Operator Performance. *Ergonomics*, 1–34.

CHAPTER 3:

The poorer the better: new evidence that low-degree automation preferentially benefits poor multi-taskers

Chapter Abstract

Background: This chapter will examine whether the impact of low-degree automation (DOA) on performance, situation awareness and workload vary with individual differences in multi-tasking ability. One conventional approach to examining automation's impact has been to compare outcomes across different DOA. Chapter 2 established an alternative approach, showing benefits can vary within a single degree of automation when accounting for operators' cognitive abilities. Chapter 3 seeks to expand this approach established in the previous chapter to new tasks demonstrating generalisability, as well as conceptually replicate findings from the previous experiment. The current chapter also attempts to detect effects for SA and workload not found in the Chapter 2 with more sensitive measures. **Methods:** One-hundred-and-ten undergraduate students completed four cognitive tasks (visuospatial tracking, attentional blink, dual task, psychological refractory period) and two conditions (manual – no automation; and low-degree automation) in an air traffic control task (ATC). A multi-tasking index was created by applying latent factors methodology to these cognitive tasks. **Results:** Linear mixed effects models indicated that the benefit of low-DOA compared to manual acceptance and hand-off performance was greater for individuals with poorer multi-tasking ability. The effectiveness of automation was found to vary within a DOA when accounting for individual's multi-tasking ability, partially replicating the previous chapters' findings. Better multi-taskers had better SA than poorer multi-taskers. Low-DOA improved conflict detection performance regardless of multi-tasking ability and reduced subjective workload across two measures. **Conclusions:** Thus low-DOA benefit was extended to a different task within the ATC task. Objective workload was not predicted by multi-tasking ability or low-degree automation.

Introduction

In the modern workplace, the need to perform multiple tasks by strategically dividing and directing attention (called 'multi–tasking'; Gutzwiller et al., 2019; Pashler, 1984) can over-load humans' limited cognitive resources (Wickens & Dixon, 2007; Wong & Seet, 2017). When this occurs, task performance typically declines, and the mental cost of performing a task, called workload, increases (Moray, 1979; Wickens, 2008). Importantly, however, the magnitude of performance degradation may not be the same for everyone. Individual differences in multi-tasking ability have been shown to modulate performance (See Chapter 2; also, Rubinstein et al., 2001), such that individuals with greater multi-tasking abilities may experience less performance degradation (Schumacher et al., 2001).

To reduce task requirements to within the limits of humans' cognitive resources (Kaber & Endsley, 2004; Saqer & Parasuraman, 2014) automation can assist with performing tasks traditionally done by humans (Parasuraman et al., 2000). Reliable automation that performs as an operator expects (Onnasch et al., 2014) can free-up cognitive resources (S. Chen et al., 2018; Kaber & Endsley, 2004) and thereby benefit performance (Kaber & Endsley, 2004; Manzey et al., 2012), and reduce workload (Parasuraman et al., 2009; Strybel et al., 2016) compared to performing a task without automation. Automation can also impact the operators' situation awareness (SA) defined as "perception and comprehension of dynamic elements in the environment and projection of their future states" (Endsley, 1988) compared to performance without automation. The nature of this impact, however, appears to depend on the type of automation. For example, SA can be improved by automation designed to support human decision making (Kaber & Endsley, 2004; Parasuraman et al., 2009), but impaired by automation that performs tasks with little human input (Onnasch et al., 2014).

Previous research has often focused on how automation modulates performance, workload and SA compared to manual performance (see Onnash et al., 2014 for review). However, prior to this thesis studies have not directly examined whether individual differences in multi-tasking could have similar effects and/or modulate automation. To address this question, Chapter 2 in this thesis was the first study that I am aware of to examine whether individual differences in multi-tasking could modulate automation's benefits to performance, workload, and SA. These outcomes were examined in a simulated Air Traffic Control (ATC) task, using a low-degree of automation (DOA) that highlighted task-relevant information. Critically, Chapter 2 found the benefit conferred by low-DOA on performance varied with multi-tasking ability. These findings are important because they indicated that the benefit of a given automated aid may not be uniform across all operators. Knowing who benefits most from automation may inform the design and implementation of more effective automation in the future.

The impact of multi-tasking ability on automation effects

In the ATC task described in Chapter 2, participants monitored multiple aircraft at varying altitudes, trajectories, and speeds, and had to: a) accept aircraft approaching their controlled airspace, b) hand-off departing aircraft, and c) detect conflicts that occurred when two aircraft violated minimum separation requirements. Low-DOA provided participants with visual cues that highlighted aircraft requiring acceptance or hand-off thereby ameliorating multi-tasking requirements by reducing the need for time-consuming visual search for incoming/outgoing aircraft. The key findings were that acceptance and hand-off performance was better when assisted by low-DOA compared to no-automation (manual condition). Moreover, this benefit was greater for poorer multi-taskers than for better multi-taskers. It was concluded that this difference arose because although automation freed-up a similar amount of cognitive resources for all participants, these resources were proportionally more beneficial for poorer multi-taskers who had lower baseline capacity to meet task requirements.

While Chapter 2 demonstrated that multi-tasking ability modulated the benefits of low-DOA to acceptance and hand-off performance as described above**,** it is not yet known whether an analogous outcome exists if automation is applied to tasks that more closely match the requirements of those tasks which are typically subject to automation in complex environments. In particular, as will be discussed below, the acceptance and hand-off task were relatively unambiguous and predictable. By contrast, the types of tasks typically subject to automation in real workplaces are more complex – meaning they require decisions under ambiguous and dynamic conditions, with less predictable outcomes, and/or more infrequent operator interventions (see Chérif et al., 2018 for review of task complexity in military operations). Examples of such task domains include uncrewed aerial vehicle operation (e.g., Rovira et al., 2007; Wong & Seet, 2017), driving autonomous vehicles (e.g., Dikman & Burns, 2016), surgery (e.g., Manzey et al., 2012), financial trading (e.g., Li & Burns, 2017), and ATC (e.g., for review see Mogford et al., 1995).

Uncovering whether multi-tasking modulates automation in these kinds of tasks is increasingly important as the need for automation in these domains continues to grow. In the ATC domain, for example, studies have pointed to the significant growth in air traffic projected for the near future, which threatens to overwhelm capacities of both human air traffic controllers and current automation (Leiden et al., 2003; Trapsilawati et al., 2017). With these points in mind, the current chapter applies low-DOA to the conflict detection task in simulated ATC as an exemplar of a type of task which is typically subject to automation and will continue to be relevant into the future.

Like many real-world tasks that are currently subject to automation, conflict detection in the current simulated ATC task requires decisions under complex conditions similar to those described above. For example, ensuring safe minimum separation standards requires projecting

aircraft trajectories into the near future. This involves judging aircrafts' relative lateral and vertical (altitude) positions within the sector and integrating temporal information to predict the future relative position based on airspeed to determine whether aircraft in close temporal proximity will cross paths (Loft et al., 2007; Xu & Rantanen, 2003). These judgements of lateral and/or vertical separation have been studied in both student (Loft et al, 2007; 2004) and expert controllers (Loft et al., 2009; Vuckovic et al., 2013; Zhang et al., 2021). Subsequent decisions must then be made to prioritize the aircraft most likely to violate separation standards as well as to appropriately intervene only on occasions that separation violations occur without intervention.

These task requirements place considerably greater strain on operator cognitive resources than aircraft acceptance or hand-off decisions. First, any aircraft could be potentially involved in a conflict, and conflicts occur at unpredictable intervals. Further, aircraft may be in close spatiotemporal proximity but not breach separation standards (i.e., a near-miss), thus requiring further observation (Bowden et al., 2021). By comparison, in the acceptance and hand-off tasks, aircraft enter or leave the simulated airspace at discrete intervals as signified by physical contact with clearly marked airspace boundary lines. Thus, acceptances and hand-offs involve less ambiguity.

As the above task analysis suggests, conflict detection may have decidedly greater reliance on cognitive abilities associated with multi-tasking than acceptances or hand-offs. Conflict detection requires active and sustained monitoring (Remington et al., 2000), and frequently switching attention between aircraft to update estimates of their relative separation ('detection requirements'; see Stankovic et al., 2011; also see Vuckovic et al., 2014). Switching also creates task interruptions that require participants to remember to come back to check on potentially conflicting aircraft at a later time (Wilson et al., 2018, 2020). By contrast, acceptances and hand-offs only require attention at one point in time – when the aircraft reaches the sector boundary. Working memory, a key component of multi-tasking (Hambrick et al., 2010), is also involved in storing and processing aircraft information. Similarly, attentional control, the ability to switch attention between tasks (J. Chen & Terrence, 2009) is required to respond to shifting cognitive and perceptual priorities involved in overseeing conflict detection across multiple potentially conflicting aircraft pairs.

The current study

The primary aim of this chapter was to investigate whether individual differences in multi-tasking ability modulated the benefit of perfectly reliable automation on conflict detection performance. To this end, low-DOA was applied to the conflict detection task in a similar manner to acceptances and hand-offs in Chapter 2, such that all pairs of aircraft that may potentially conflict (i.e., 10 conflict pairs plus six near-miss pairs which were spatiotemporally proximal but did not violate separation standards) were highlighted red once both aircraft in the pair were

accepted. The highlighting remained visible until either: a conflict occurred, a conflict was detected and resolved (i.e., averted), or, in the case of near miss pairs, the aircraft passed their closest point of separation without conflicting. In addition to conflict detection, participants completed the same acceptance and hand-off tasks as in Chapter 2 with low-DOA and manually. Subjective and objective workload, and situation awareness were also measured. To assess multi-tasking ability, participants completed four cognitive tasks prior to the ATC task. As in Chapter 2, performance on these tasks were concatenated using latent factors analysis to provide a single multi-tasking ability estimate.

Previous studies have suggested that different types of automation can broadly benefit conflict detection. For example, Trapsilawati et al. (2017) found when low-DOA, which alerted operators of potentially conflicting aircraft and recommended intervention actions, was applied to conflict detection, air traffic controllers' workload was reduced and conflict detection performance improved. Similarly, a study by Vuckovic et al. (2013) with air traffic controllers who completed conflict detection scenarios with radar only, and with radar and automation (called the 'multi-conflict display' which highlighted aircraft on the same or different altitudes) found performance to detect actual or potential conflicts was better with automation, particularly when many aircraft were on the screen.

Finally, Galster et al. (2001) found air traffic controllers' conflict detection accuracy improved and response time decreased with high-DOA under high-traffic conditions, however subjective workload also increased. High-DOA may have helped controllers keep up with traffic volume but was less effective at supporting decision making. However, none of these studies examined whether the magnitude of these benefits varied based on individual differences in multi-tasking ability.

In the current chapter, several predictions were made. Given previous findings of reduced workload and improved performance when low-DOA was applied to conflict detection (Trapsilawati et al., 2017), and the fact that that low-DOA benefited acceptances and hand-offs in Chapter 2, it was predicted that automation would benefit conflict detection performance. As described earlier, conflict detection has significant multi-tasking requirements. Therefore, greater multi-tasking ability was also expected to benefit conflict detection. Finally, as was found in Chapter 2 for acceptances and hand-offs, it was also expected that multi-tasking ability would modulate the benefit of low-DOA for conflict detection performance.

There were several possible outcomes as to how multi-tasking may interact with low-DOA to predict conflict detection. One outcome is that poorer multi-taskers may experience greater benefits when low-DOA assists conflict detection, compared to low-DOA assisting acceptances and hand-offs in Chapter 2. If this was the case, then multi-tasking may modulate the benefit of low-DOA on conflict detection to a *greater extent* than for acceptances and hand-offs. Alternatively, as conflict detection is more complex and ambiguous, the benefit received from

low-DOA may be similar for all participants regardless of their multi-tasking ability. Thus, multi-tasking may modulate the benefit of low-DOA to a *lesser extent* than acceptances and hand-offs.

ATC performance replication and methodological changes to SA and workload measures

The current methods were similar to Chapter 2 which provided an opportunity to replicate the key findings. Should the interaction between multi-tasking and low-DOA be replicated across tasks (i.e., acceptances and hand-offs) or conditions (i.e., experiments), this would demonstrate strong generalisability of the underlying assumption of multi-tasking driving automation effects across a larger sample of individual differences. Replicating Chapter 2, multi-tasking ability was expected to modulate the benefit of low-DOA for acceptance and hand-off performance.

This chapter also took the opportunity offered by partial replication to re-examine SA by improving the queries used in Situation Present Assessment Method (SPAM; see Durso et al., 1998). To this end, the number of queries related to acceptance, hand-off, and conflict detection tasks were more closely balanced, and examined item wording and item reliability to improve query properties. These changes were intended to increase the sensitivity (i.e., likelihood of finding effect of experimental manipulation) of the SA measure (see Methods section) in order to maximize the potential to detect any effect of low-DOA on SA (which was not found in Chapter 2). Because of these improvements, multi-tasking ability and low-DOA were predicted to modulate SA, and this chapter expected to find an interaction between these factors such that the benefit of low-DOA decreases as multi-tasking ability increases.

This chapter also employed the NASA-TLX (Hart & Staveland, 1988), a widely used measure of subjective workload (for meta-analysis see Grier, 2015) that is reliable and well validated (Hart & Wickens, 2006). There were two reasons for including the NASA-TLX here. First, it could potentially provide converging evidence for the single-item post-scenario measure of subjective workload used in Chapter 2. Second, the NASA-TLX includes subscales that tap different facets of workload. Some of these sub-scales are potentially more relevant to multi-tasking such as 'mental workload', 'temporal workload' and 'effort'. Subjective workload was expected to be lower in the low-DOA condition compared to manual performance as in Chapter 2. Multi-tasking may predict subjective workload with the inclusion of the NASA-TLX which examines multi-tasking-relevant processes (e.g., mental, and temporal workload). Objective workload was again examined using latency of accepting SPAM queries to investigate the benefit of low-DOA, noting that in this chapter all aspects of the ATC task (acceptances, hand-offs, and conflict detection) were subject to automation which may increase the possibility of finding lowered objective workload in the low-DOA condition.

Finally, this chapter also employed a new multi-tasking paradigm designed to improve the structure of the latent factor. The latent factor model in the previous chapter was derived from two tasks (the Psychological Refractory Period and Dual task) administered to participants. While

this yielded a factor that explained an acceptable degree of variance, it has been suggested that a three-task latent factor is preferable to the two-task factor to decrease the risk of creating an under-defined model (Draheim et al., 2019). To this end, this chapter administered a fourth task – the visuospatial tracking which – along with the PRP, dual task and AB tasks used in the previous chapter. The visuospatial tracking task (VST: Bender et al., 2018) required a high degree of dual-tasking, and the ability to rapidly switch between tasks over a relatively short time span. The VST is highly representative of multi-tasking as it involves performing a go-/no-go task while also tracking a circle moving randomly on the task screen.

Methods

Participants

University undergraduate students (N = 120) were recruited from a psychology research participation pool. Participants provided informed consent and received AUD$10 and partial course credit. Data from ten participants were omitted because of incomplete data (see below) leaving a sample of 110 participants (37 males, 72 females, one not specified, M age = 23.20, range = 18 – 58). This research complied with the tenets of the Declaration of Helsinki and this project was approved by the Human Research Ethics Committee of The University of Western Australia.

Measures

The operationalisation of multi-tasking using latent factors methodology was a key strength of Chapter 2. This method benefits from combining shared variance reflecting the multi-faceted nature of cognition rather than using a single task to measure cognition as is more commonly done. One limitation of Chapter 2 however was that the latent factor was based on only two variables as the third cognitive task, the Attentional Blink (AB) was excluded from the factor because it did not load with the Psychological Refractory Period (PRP) and dual response selection (Dual task) tasks. A three-task latent factor would therefore provide a more robust representation of multi-tasking ability. The visuospatial tracking task (VST) was included in this chapter to maximize the chance of creating a three-task latent factor. The AB was retained as it is theoretically relevant, tapping into rapid task switching and encoding processes involved in multi-tasking, and has previously found to load on a latent factor with the Dual task (Bender et al., 2017).

By way of a brief recap of the cognitive tasks used for the multi-tasking latent factor, the AB represented millisecond processing speed decrements due to the close temporal proximity of two visual stimuli (Chun & Potter, 1995). The PRP assessed limitations to participants' ability to process two stimuli simultaneously (Van Selst et al., 1999). The dual response selection task estimated the cost of concurrently performing two tasks compared to performing a single task

(Dux et al., 2009). All measures were the same as Chapter 2, with the following tasks added or changed.

Visuospatial Tracking Task (Bender et al., 2018): The VST, based on a technique by Anguera et al. (2013), assesses the performance cost of concurrently tracking a moving target (white circle) using the mouse and completing a centrally-presented go/no-go task consisting of a sequence of shapes that included targets, defined by a shape and color combination (e.g., red triangle) amongst a set of distractors which were other shapes and different colors to the target (e.g., blue triangle and red squares).

The task contained two phases. First, a thresholding phase determined task difficulty based on an individual's performance. A staircase procedure was used to increase or decrease the level of difficulty by one step (min = 0, max = 40 steps) for each trial type (four single-tracking and four single shape detection trials) to determine the difficulty level at which participants' accuracy was 80% (+/-5%). The difficulty level associated with performance closest to 80% was applied in the experimental phase which contained 24 trials across three conditions (8 trials × 3 conditions); single-tracking (30 seconds), single shape detection (go-no-go; 60 seconds) and combined 'dual task' condition (60 seconds) in which participants performed both tasks concurrently. Task instructions to participants can be found in supplementary materials.

Performance - Air Traffic Control Task (Fothergill et al., 2009): The same ATC task was used as Chapter 2. Participants completed two counterbalanced ATC conditions, with one critical difference from Chapter 2 – the addition of low-DOA for the conflict detection task. Aircraft that potentially could conflict turned red once both aircraft were accepted and remained so until the conflict was resolved, or in the case of near misses, the aircraft passed their closest point. This was applied to the 10 pairs of actual conflicting aircraft plus the six pairs of near-miss aircraft which did not breach minimum separation standards but passed each other just outside minimum distances. In the manual condition, aircraft pairs were not highlighted. As in Chapter 2 in the low-DOA condition, aircraft requiring 'acceptance' and 'hand-off' during the 20-second response window flashed and changed colour (acceptances = blue, hand-offs = orange). In the manual condition, all aircraft remained static green colour with no flashing or colour change.

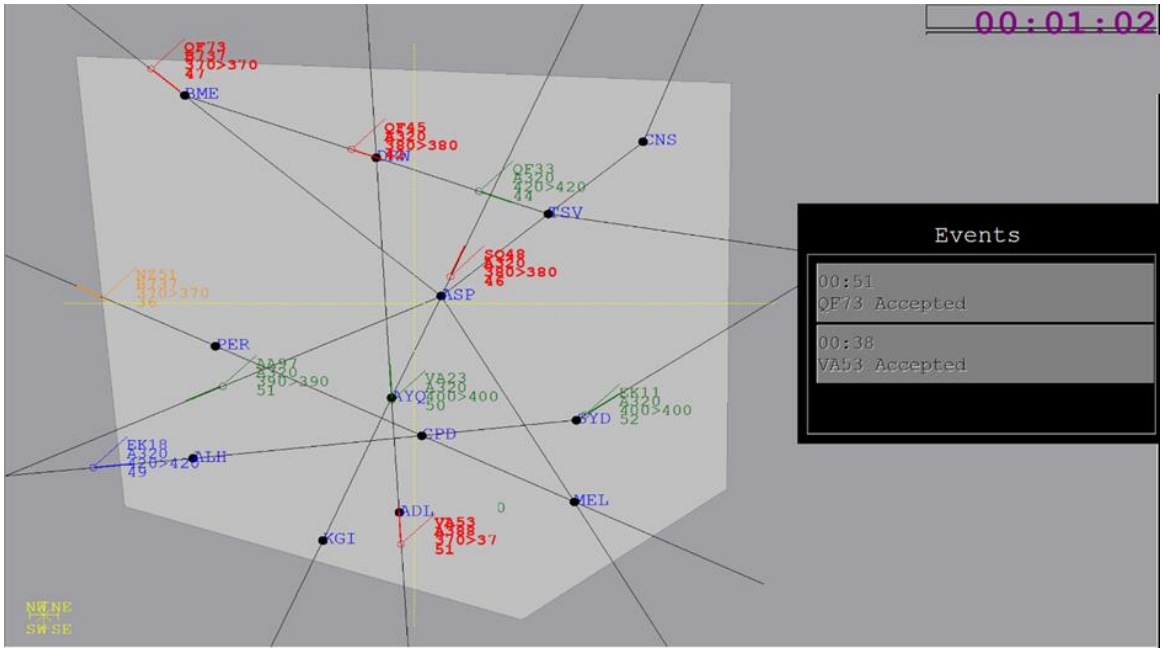


Figure 3.1. Air traffic control sector display. In the low-DOA condition (shown here), aircraft represented by circles flashed blue when they approached the participants' sector (light grey polygon), indicating they required acceptance (i.e., EK18) and orange indicating they required hand-off (i.e., NZ51). Additionally, pairs of aircraft (i.e., QF45 and SQ48 both at 38,000ft; also, QF75 and VA53 both at 37,000ft) turn red if they potentially conflicted (includes near misses) in the future. In manual condition, all aircraft remained green. Actions performed by the participant were logged in the 'Events' box on the right of screen. Black lines indicate flight paths.

Situation Awareness: Situation awareness was measured using a modified Situation Present Awareness Method (SPAM; Durso et al., 2004). There were again 18 queries per condition appearing at 2–3-minute intervals. As in Chapter 2 participants received a visual prompt ('Ready for Question?') before each query and were instructed to click on this prompt as soon as their workload permitted. SA query was presented with four response options at the top right of the screen. The ATC task was then paused but not blanked. Participants were instructed to respond as accurately and quickly as possible.

Following the results in Chapter 2 the opportunity was taken to systematically evaluate the SA queries in terms of clarity, content validity (i.e., how well the queries fit the scenario in terms of timing of the query and the events they assessed), and scale reliability if items were removed. Queries were also ensured to equally assess acceptances, hand-offs, and conflict detection information. These minor changes to some SA queries may increase the sensitivity of the measure and thereby the likelihood of detecting an effect of low-DOA on SA.

Based on Chapter 2 data, items with an accuracy lower than 1.5x the standard deviation below the group mean accuracy (a lower criterion of 70%) were flagged. This led to the examination of items such as "In which quadrant will the next loss of separation take place within the next 30 secs if no action will be taken?" (query number 3 of scenario 1 with 60% accuracy) and "How many aircraft needed accepting within the last 30 secs in the SW quadrant?" (query

number 7 of scenario 1 with 65% accuracy). The first of these examples was found to have several ambiguities in terms of confusing wording, and timing as participants may have already detected the next conflict and intervened, or may not have, in which case the query cues them to the existence of the conflict. This query also was associated with low average in each instance it was used (72% and 73% in the second scenario). For consistency, all instances of this query were replaced. In line with best practice guides set out by Endsley (2021) that queries should be clear and consistently scored, and a systematic analysis of SA requirements for the domain (e.g., ATC, submarine, driving) should be employed to avoid "the development of queries that may in fact have little to do with the operator's SA requirements" (Endsley, 2021, p. 19), ambiguous, vague or inconsistently worded queries were removed and replaced with more specific phrasing.

Secondly, scale reliability was assessed, from the Chapter 2 data and an alpha of .76 was found for both sets of SA queries. The reliability of individual queries was then assessed by examining the scale alpha if each query was removed. Three queries per set were flagged as their removed improved the scale reliability close to an alpha of .80. Some of these queries with low reliability overlapped with ones identified as low accuracy.

Thirdly, queries were also found to address conflict detection SA more often than other aspects of the task (acceptances and hand-offs). New queries were created to ensure an equal number of queries across tasks (acceptances and hand-offs, conflict detection) and SA levels (past, present, future) to ensure all aspects of the task were examined equally so that potentially different effects of low-DOA on these ATC tasks would not bias the SA result. Statistical examination of queries used in Chapter 2 can be found in supplementary materials Tables 4 and 5. The new queries used in the subsequent Chapters is shown in supplementary materials Table 6.

Workload: At the end of each ATC condition, participants completed an online version of the NASA-TLX (Hart & Staveland, 1988); that first required participants to rate six workload subscales (mental, physical, and temporal demand, performance, effort, frustration) on a 21-point scale from 0 (low) to 100 (high). Each subscale was defined and included synonyms. Participants choose between 15 pairs of subscales (all combinations of the six subscales) to indicate which they consider the 'more important contributor' to their workload during the task (e.g., 'Performance' OR 'Frustration'). Each scale's rating was then weighted by the number of times it was selected in the combinations (zero to five) to produce a weighted subscale score. A global workload score ranging from 0 (low) to 100 (high) was calculated as the weighted subscale scores averaged (Battiste & Bortolussi, 1988).

Other Measures: Before completing the cognitive tasks, participants completed a demographics questionnaire and a 9-item subscale from the Raven's Standard Progressive Matrices to examine non-verbal intelligence (RSPM. Bilker et al., 2012), and other personality questionnaires were included for potential exploratory analysis outside the scope of this thesis. They assessed traits including narcissism using the 20-item Narcissism Scale (Derry et al., 2017),

self-esteem, using the 10-item Self-esteem questionnaire (Rosenberg, 1965), and extraversion, conscientiousness and honesty using the modified BFI - 6 personality inventory (Thalmayer et al, 2011). Additional questionnaires measured state anger using the 15-item State Anger STAXI-II (Spielberger et al., 1999), shame using the 15-item State Shame Scale (Marschall, 1996), and self-esteem using the 7-item State Self-esteem Scale (Heatherton & Polivy, 1991). After each ATC condition, participants completed a 9-item measure of boredom (Fahlman et al., 2013) and fatigue (Chalder et al., 1993), the mental toughness scale (Gucciardi et al., 2015), and a questionnaire that asked participants to rank the relative importance of hand-offs, acceptances, and conflict detection. These measures were not analysed for the purpose of this thesis.

Procedure

All tasks were completed on a Windows PC, with a BENQ 24$^{\text{-inch}}$ monitor, providing a screen resolution of 1920 × 1080 running at 100Hz refresh rate. Participants first completed the demographics questionnaire. Then the cognitive tasks were completed in counterbalanced order. Participants then were presented with the 20-minute training video and comprehension questions described in Chapter 2 outlining the manual and automated ATC tasks, a 20-minute manual practice condition, and two 30-minute experimental conditions in counterbalanced order. Lastly participants completed the post-condition subjective workload questionnaire and other questionnaires outlined above that were not analysed in this thesis. After the testing session, participants were debriefed, given the opportunity to ask additional questions, and received remuneration. Testing sessions took 3.5 hours.

Results

Data cleaning

For all cognitive tasks requiring a speeded response (including the ATC task), mean response time (RTs) were calculated using correct responses trials with RTs longer than 150ms (except conflict detection) or less than 3 SD above each participant's mean RT to exclude outliers. This was the only data cleaning applied to the Dual task (Table 3.3). Data cleaning resulted in 1.20% of Dual task data being removed. Specific data cleaning procedures for the other cognitive tasks in addition to this are described below.

PRP Task: In addition to the data cleaning described above, PRP calculations omitted RTs if there was less than 50ms between keyboard responses (Tombu & Jolicœur, 2005). Mean RTs were calculated separately for each stimulus type (visual or auditory: PRP; Table 3.2). Data cleaning resulted in 15.59% of PRP trials being removed.

AB Task: Accuracy for T2 was calculated only for trials in which the first target (T1) was correctly identified (Raymond et al., 1992). Mean T1 and T2|T1 accuracy were calculated separately as a function of lag (Table 3.1). No AB data was removed.

VST Task: VST performance was measured in mean accuracy for each condition and tasks in the dual condition separately (i.e., tracking mean and shape detection performance). Single-task accuracy was checked to ensure it was above 80% for each participant as intended by the staircase procedure. No participant's results were below this minimum threshold. Dual cost (called 'tracking cost' and 'shape cost') was also calculated as the difference between mean tracking dual task and single-task accuracy (Table 3.4).

ATC task: Mean response time (RTs) were calculated using correct responses trials with RTs greater than 150ms (except conflict detection) or less than 3 SD above each participant's mean RT to exclude outliers. This data cleaning removed 3.17% of acceptance, 3.30% of hand-offs, and no conflict trials. Adjusted RTs were calculated separately for acceptances, hand-offs, and conflict detection trials by dividing mean RTs by the corresponding mean accuracy, which combines accuracy and RT into a single composite measure accounting for speed-accuracy trade-offs (Liesefeld & Jancyzk, 2019; Visser et al., 2015).

SA: RTs on correctly answered SA queries were used to calculate an adjusted SPAM RT reflecting the quality of SA. Data cleaning removed 1.45% of SA trials.

Workload: Objective Workload RTs were calculated only on trials in which the "Ready" prompt was responded to before the deadline, and the SA query was correctly answered. For the single-item subjective workload measure, data from one participant was omitted due to incomplete responses. Additionally, global scores from the NASA-TLX were calculated in each condition. The NASA-TLX had a smaller sample of 79 as a glitch in the administering of the measure meant data was not collected from the first 40 participants. Data from two additional participants could not be used due to missing responses. This technical issue removed 1.45% of objective workload, subjective workload single-item (1.26%), and NASA-TLX (2.53%).

Table 3.1. Mean percentage of targets correctly identified in the AB task with standard deviation in parentheses.

	T1	T2|T1
Lag 1	81.90 (16.33)	85.54 (14.71)
Lag 3	86.56 (15.13)	45.77 (22.00)
Lag 8	86.29 (13.89)	80.32 (16.40)

Table 3.2. Mean percentage of targets correctly identified and mean response time (in milliseconds) in the PRP task, separated by target type, with standard deviation in parentheses.

Inter-target interval	Sound trials		Visual trials	
	Accuracy	RT	Accuracy	RT
200ms	98.71 (9.45)	2019 (744)	98.50 (9.43)	1639 (737)
1000ms	97.71 (9.44)	1459 (555)	97.46 (9.75)	1696 (776)

Table 3.3. Mean percentage of targets correctly identified and mean response time (in milliseconds) in the Dual Response Selection Task, separated by target type, with standard deviation in parentheses.

Trial Type	Shape		Sound	
	Accuracy	RT	Accuracy	RT
Single task	92.95 (10.84)	767 (152)	92.32 (10.80)	852 (159)
Dual task	90.05 (11.60)	953 (191)	90.00 (11.60)	1095 (196)

Table 3.4. Percent of trials correctly identified in the Visuospatial Tracking Task with standard deviation in parentheses. Performance cost reflects the difference in accuracy between single and dual conditions

Trial Type	Shape Go/No-Go	Disk Tracking
	Accuracy	Accuracy
Single task	80.33 (14.07)	82.04 (9.32)
Dual task	64.54 (12.27)	79.03 (9.36)
Performance Cost	15.76 (7.05)	3.01 (3.42)

Table 3.5. Mean and Standard deviation of ATC performance and RT (seconds). Standard Deviations in Parantheses. Note that lower adjusted RT = better performance/SA

	Manual		Low-DOA	
Task	Accuracy	RT	Accuracy	RT
Acceptance	94.78 (14.11)	3.91 (1.42)	99.48 (9.93)	2.30 (0.82)
Hand-off	95.95 (12.83)	3.59 (1.41)	99.69 (9.97)	2.59 (0.88)
Conflict detection	95.95 (12.31)	97.67 (30.01)	96.72 (11.31)	82.24 (35.38)
SPAM	89.95 (11.84)	16.12 (6.07)	88.60 (12.26)	15.63 (5.53)

Latent factor analysis

As previously, a preliminary exploratory factor analysis was conducted with the PRP (Short PRP RT) and dual (Dual RT) variables used in Chapter 2, AB magnitude variable described in Chapter 2, and the new VST variables (tracking dual cost, tracking dual mean, shape dual cost and shape dual mean). It was found none of the VST variable loaded acceptably (highest loading = 0.15) with the PRP, Dual and AB variables. The VST task was then eliminated from further factor analysis. The AB magnitude in this chapter was found to load adequately with the PRP and Dual variables used in Chapter 2, and thus the AB was included in the factor in this Chapter. The final EFA was conducted using Principal Axis Factoring extraction and Direct Oblimin rotation with the PRP, dual and the AB magnitude variables. Individually, these tasks represent the more difficult condition within each task, thus reflecting multi-tasking ability. The loadings were adequate (Dual = 0.77, uniqueness = 0.40; PRP = 0.80, uniqueness = 0.35; AB = 0.49, uniqueness = 0.75). The factor explained 49.6% of the variance. A multi-tasking factor score was created by for each individual by saving the Bartlett scores from the factor analysis which isolate shared variance on a factor across the tasks included (DiStefano et al., 2009). Scores were transformed (multiplied by -1) so that higher scores represented better multi-tasking ability (Bartholomew et al., 2009).

Linear mixed models

This Chapter replicates the analysis approach used in Chapter 2. To examine if performance differed as a function of multi-tasking ability and the presence of low-DOA, this chapter conducted a series of linear mixed effects models (LMM). Data was entered in a nested form to account for the within-subject design. Condition was entered with two levels – automated or manual as a fixed factor and multi-tasking entered as a fixed continuous factor. An interaction term was entered as a fixed effect to examine the interaction of multi-tasking and condition. Participant number was entered as a random effect (intercept) and condition as a random effect (slope) to control for within-subject variability differing across conditions.

An LMM was fitted for each dependent variable. Model significance was assessed using a Chi-square test to compare to a null model which included only the random effects and the intercept (see supplementary materials Table 7). Assumptions of linearity, absence of autocorrelation, absence of influential cases and multicollinearity were satisfied (see supplementary materials Assumption plots). Homoscedasticity was at an acceptable level (i.e., evenly distributed) for all models. The assumption of normality of residuals was violated in all models, however LMM have been found to be robust to violations of this assumption (Knief & Forstmeier, 2021; see review by Schielzeth et al., 2020).

Models were initially conducted with RSPM, previous experience with ATC, and counterbalance order as covariates to control for potential effects of strategy, practice effects,

intelligence, and task order. RSPM was included to ensure the variance explained by multi-tasking ability was not due to its overlap with non-verbal intelligence. Past ATC experience was a significant covariate of SA, and intelligence was a significant covariate of acceptances and hand-offs (see supplementary materials Table 8). However, it should be noted the same pattern of significant and non-significant effects that are reported below remain so even if these covariates are included in the models. Thus, for brevity, these covariates are not discussed further.

Standardised betas are presented in tables to allow comparison of effect sizes across studies and measures. Negative betas for multi-tasking indicate that adjusted RT decreased as the model moved across the predictor (i.e., from left to right on the X axis – multi-tasking scores). Likewise negative betas for condition indicate adjusted RTs decreased as the model moved from 0 (manual) to 1 (low-DOA) conditions. Decreased adjusted RTs reflect improved performance (i.e., faster and/or more accurate).

ATC Performance: Accuracy and RT data for the ATC tasks and SPAM measure is presented in Table 3.5. Standardised coefficients and model fit indexes are presented in Table 3.6. For conflict detection, graphical representation suggests a multi-tasking effect was present in the manual condition, but not in the low-DOA condition where participants performed similarly regardless of cognitive ability (see Figure 3.2). However, there was no multi-tasking effect and no interaction. Only a significant condition effect was found, indicating overall conflict detection performance was better in the automated condition than in the manual condition.

For acceptances and hand-offs, the significant condition effect again indicated overall superior performance in the automated condition compared to the manual condition. In addition, a significant effect of multi-tasking was found which indicated that participants with higher multi-tasking scores performed better overall. A significant interaction was also found between multi-tasking and condition for acceptances and hand-offs indicating performance improvements in the low-DOA condition compared to the manual increased as multi-tasking scores decreased (see Figure 3.2). These two findings replicate results obtained in Chapter 2.

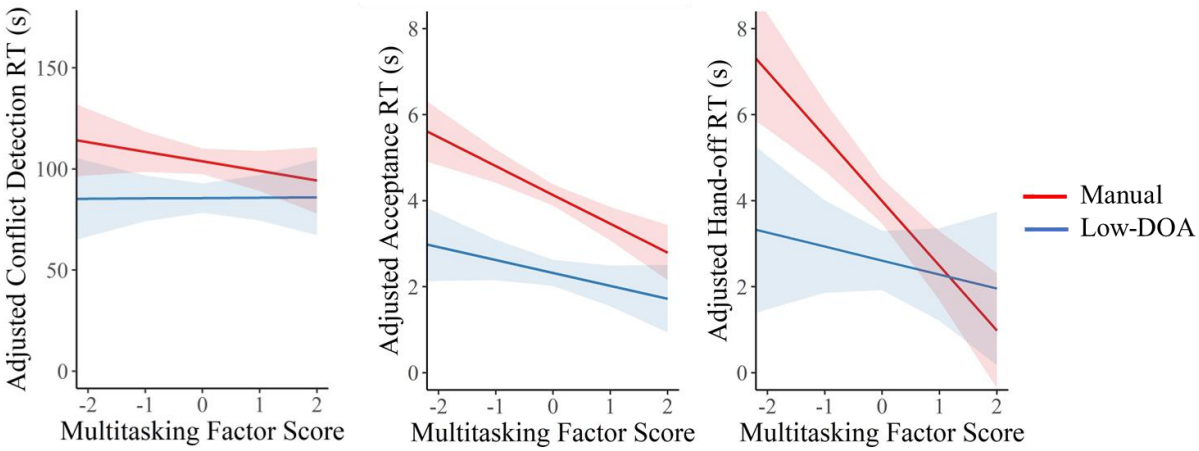


Figure 3.2. Conflict Detection (left), Acceptances (centre) and Hand-offs (right) LMM slope and intercept. Performance measured in adjusted response time (seconds) against multi-tasking score (standardised). Lower adjusted RT reflects better performance. Higher multi-tasking scores reflect better multi-tasking ability. Shaded area represents 95% CI.

Table 3.6. Linear mixed effects standardised coefficient estimates, standard error in parentheses and model fit summaries

	Dependent variables				
Parameter	Acceptance Adjusted RT	Hand-off Adjusted RT	Conflict Detection Adjusted RT	SA Query Adjusted RT	Objective workload (single item)
Condition	-0.33 ***	-0.21***	-0.14 ***	-0.02	-0.004
	(0.10)	(0.24)	(2.96)	(0.34)	(0.07)
Multi-tasking	-0.14 ***	-0.23 ***	-0.03	-0.25 ***	-0.05
	(0.15)	(0.31)	(3.85)	(0.54)	(0.08)
Condition ×	0.06 **	0.16***	0.03	0.01	0.02
Multi-tasking	(0.12)	(0.28)	(3.52)	(0.41)	(0.08)
Observations	9387	9536	2119	3441	3396
Log Likelihood	-21594.30	-23604.90	-11607.30	-12186.20	-6435.60
AIC	43202.54	47223.80	23228.50	24386.50	12885.20
BIC	43252.60	47273.90	23268.10	24429.50	12928.10
R2 (conditional)	0.32	0.53	0.24	0.28	0.14
R2 (marginal)	0.12	0.07	0.02	0.06	0.003

Note: *** $p < .001$, ** $p < .01$, * $p < .05$. The estimates represent standardised coefficient estimates for fixed effects in the linear mixed model on the trial level specifying a random intercept (unstandardised) and slope per participant. P-values were computed using the Wald approximation. Observations are the number of data points (trials) analyzed. Log likelihood is a measure of fit describing how well a parameter explains the observed data. AIC is Akaike Information Criterion, an extimator of prediction error. BIC is Bayesian Information Criterion, a measure for model comparison or selection. R2 conditional is a measure of model fit, indicating the percent of variance explained by the overall model. R2 marginal is the percent of variance explained by the fixed effects in isolation.

Situation Awareness: An effect of multi-tasking was found that indicated participants with higher multi-tasking scores had better SA (see Table 3.6 for standardised coefficients). There was no condition or interaction effects (see Figure 3.3). This finding replicates Chapter 2.

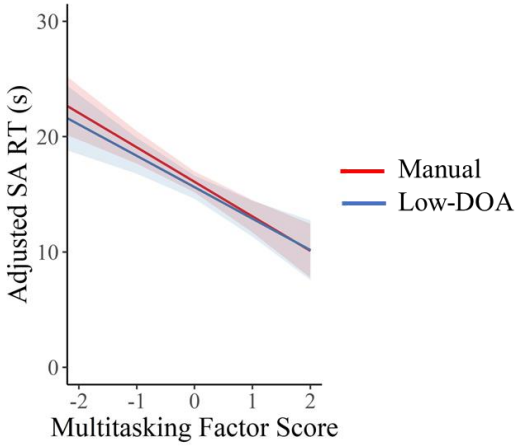


Figure 3.3. Situation Awareness LMM slope and intercept. Performance measured in adjusted response time (seconds) against multi-tasking score (normed). Black line represents manual condition and grey line represents low-DOA condition. Lower adjusted RT reflects better performance. Higher multi-tasking scores reflect better multi-tasking ability. Shaded area represents 95% CI.

Objective Workload: No effect of multi-tasking was found (see Table 3.6). There was also no effect of condition or interaction (see Figure 3.4). This differs from Chapter 2 where a small multi-tasking effect was observed.

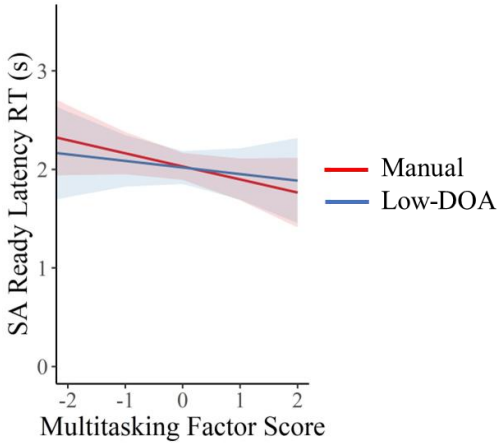


Figure 3.4. Objective workload measured by response latency to SPAM 'ready' prompt (measured in seconds) LMM slope and intercept. Black line represents manual condition and grey line represents low-DOA condition. Lower adjusted RT reflects better performance. Higher multi-tasking scores reflect better multi-tasking ability. Shaded area represents 95% CI.

Subjective Workload. Multi-tasking scores were divided into three approximately equal groups as the single-item was not suitable for LMM as exactly two data-points per participant does not allow modelling of individual random slopes. Subjective workload measure was analysed using a 2 (automated versus manual) × 3 (multi-tasking ability: low, medium, high) ANOVA which found an effect of condition $F(1,105) = 116.76$, $p < .001$, $\eta^2 = 0.25$, with workload rated lower in the low-DOA (M = 3.65, SD = 1.55) condition compared to the manual condition (M = 5.19, SD = 1.10; see Figure 3.5). There was no multi-tasking effect or interaction ($F < 2.45$, $p > .93$, $\eta^2 > .02$).

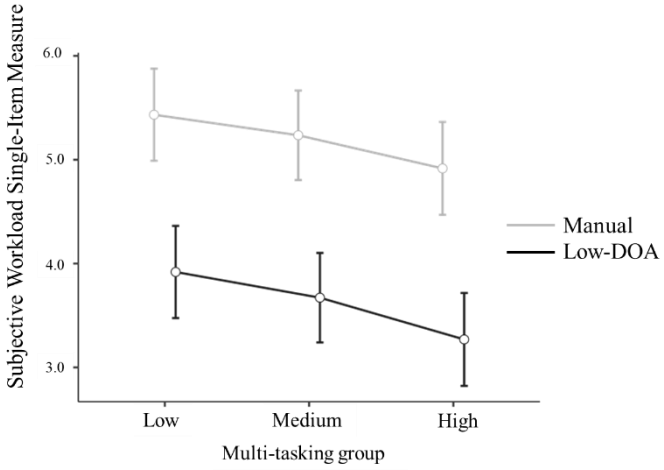


Figure 3.5. Subjective workload rating item measure across three levels of multi-tasking ability and two conditions of the ATC. Error bars represent 95% confidence intervals around the estimate of marginal means (between subjects).

The NASA-TLX global workload measure was also analysed with a 2x3 ANOVA with multi-tasking divided into the same three groups. This yielded an effect of condition $F(1,73) = 37.68$, $p < .001$, $\eta^2 = 0.15$, with workload rated lower in the low-DOA than in the manual condition. There was no multi-tasking effect or interaction ($F < 0.73$, $p > .48$, $\eta^2 > .004$). Analysis was similarly conducted on each of the sub-components of the NASA-TLX. Of these a significant effect of condition was found for effort $F(1,73) = 25.99$, $p < .001$, $\eta^2 = 0.10$; mental demand $F(1,73) = 48.53$, $p < .001$, $\eta^2 = 0.16$; and temporal demand $F(1,73) = 41.18$, $p < .001$, $\eta^2 = 0.09$, each indicating higher workload in the manual condition (see Figure 3.6 and Table 3.7 for mean and SD by condition). For each of the other sub-components there was no significant effect of condition ($F < 2.20$, $p > .19$, $\eta^2 > .009$). There was no multi-tasking effect or interaction for any sub-components ($F < 2.60$, $p > .07$, $\eta^2 > .004$)

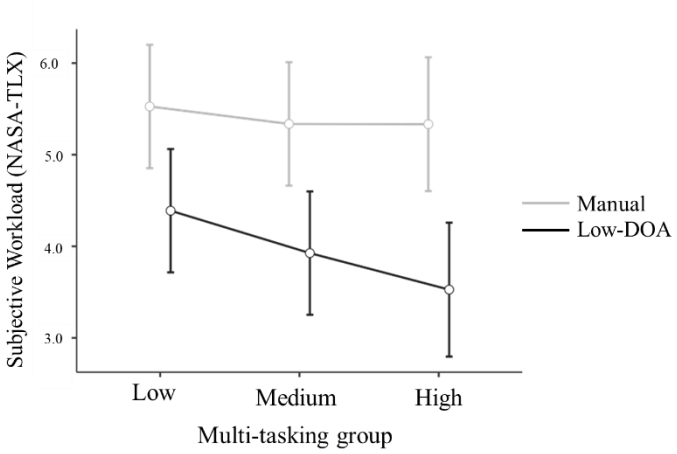


Figure 3.6. Subjective workload measure (NASA-TLX) across three levels of multi-tasking ability and two conditions of the ATC. Error bars represent 95% confidence intervals around the estimate of marginal means (between subjects).

Table 3.7. NASA-TLX Component mean and standard deviations across manual and low-DOA conditions

	Manual		Low-DOA	
	M	SD	M	SD
Effort	209	114	139	104
Mental Demand	269	125	166	122
Temporal Demand	162	111	101	93.4
Physical Demand	3.72	15.7	7.95	25.2
Performance	135	114	158	125
Frustration	21.8	61.8	23.9	68.2
Workload Total	53.5	15.6	39.9	19.4

Discussion

This thesis's aim is to examine how individual differences in cognitive ability modulate the impact of automation on performance, SA, and workload outcomes. To this end, this chapter applied automation to conflict detection to examine the potential interaction between multi-tasking and low-DOA on a more complex task of the type often subject to automation in complex work environments. The same methodology was as in Chapter 2 which provided the opportunity for a conceptual replication of the findings of Chapter 2 regarding acceptance and hand-off tasks. SA was re-examined with the inclusion of more sensitive queries to increase the possibility of detecting a low-DOA effect on SA, and the widely used NASA-TLX was added to provide converging evidence to potentially validate subjective workload effects found in Chapter 2. Objective workload was examined again for consistency across chapters.

Results showed that conflict detection performance benefitted from low-DOA as expected. However, the magnitude of this benefit did not significantly vary with individuals' multi-tasking ability. In contrast, this Chapter did find that poorer multi-taskers had the greatest benefit from low-DOA for acceptances and hand-offs, replicating findings in Chapter 2. As multi-tasking ability increased, SA increased as expected, but the low-DOA applied to all ATC tasks did not affect SA. Subjective workload was lower in the low-DOA condition as expected. Additionally, multi-tasking ability did not influence subjective workload on either subjective workload measure, and objective workload was not affected by low-DOA or multi-tasking. As expected, the low-DOA benefit did not vary with multi-tasking ability as neither of these main effects were detected.

Low-DOA benefits conflict detection

Conflict detection is a complex task with considerable multi-tasking requirements closely matching those of tasks which are automated in the real world. It requires identification of aircraft pairs (i.e., visual search), monitoring aircraft trajectories and deciding if aircrafts would conflict before performing appropriate actions to intervene. Conflict detection performance was improved when low-DOA was provided that highlighted potentially conflicting pairs of aircraft. This finding was expected given automation can free-up cognitive resources to benefit performance (see Chapter 2; Kaber & Endsley, 2004; Manzey et al., 2012). Interestingly, the benefits of low-DOA to conflict detection task were less than the benefits to acceptances and hand-off tasks which had poorer multi-tasking requirements (i.e., reduction in adjusted RT with automation: conflict detection: $\beta = -0.14$, acceptances $\beta = -0.33$, hand-offs $\beta = -0.21$).

The smaller benefit of low-DOA for conflict detection may be because automation provided less effective assistance relative to greater manual task requirements. Essentially, low-DOA reduces the detection requirements for acceptances, hand-offs, and conflict detection as it is less effortful for participants to scan and locate aircraft and monitor them until action is required. These tasks also have decision making requirements; however, they differ in the nature of the decisions. Acceptances and hand-offs were simple and required a binary decision of when aircraft reached the boundary (i.e., a predictable location). However, conflict detection was more difficult and involved uncertainty. For instance, not all aircraft at the same altitude on intersecting trajectories conflicted, and conflicts could occur anywhere inside the sector, often simultaneously. Thus, most of the difficulty in conflict detection and its critical difference to acceptances and hand-offs lies in the decision-making requirements which low-DOA may not have assisted as much as it did for detection requirements. This may be because low-DOA assists perceptual processes rather than decision making (which would be assisted by higher-DOA). In sum, the conflict detection finding suggests low-DOA can benefit more complex tasks, replicating previous

benefits found in Chapter 2 for the simpler acceptance and hand-off tasks, although to a lesser degree.

The above points notwithstanding, the finding that multi-tasking ability did not predict conflict detection performance was unexpected given the ATC literature broadly supports the high multi-tasking requirements of this task (Falkland & Wiggins 2019; Jipp & Ackerman, 2016; Luciani et al., 2020). Previously, Chapter 2 found multi-tasking ability predicted performance in all ATC tasks. As noted above, low-DOA reduced the visual search (i.e., detection) requirement for conflict detection, potentially freeing-up cognitive resources particularly for poorer multi-taskers as evidenced by their improvement from manual. Therefore, low-DOA likely equated performance for all multi-taskers by reducing detection requirements. It is likely that better multi-taskers may not have performed better with low-DOA because their detection performance could not be improved. Differences between poorer and better multi-taskers may have persisted in the decision-making requirements – however this aspect of the task was not facilitated by the provision of low-DOA. The current chapter did not test for individual differences in decision making which may show different interactions with low-DOA.

Replication of low-DOA interaction with multi-tasking ability for task performance

The current chapter replicated the overall pattern of results reported in Chapter 2 for acceptances and hand-offs. Low-DOA was expected to interact with multi-tasking because automation is intended to reduce the task load on cognitive resources required for multi-tasking. Operators with poorer multi-tasking ability benefitted proportionally more from low-DOA that assisted the multi-tasking associated with aircraft acceptance and hand-off than an operator with more multi-tasking ability. This finding from Chapter 2 was replicated here, further demonstrating that individuals can benefit differentially from the same automation, which has practical importance for future automation implementation. Additionally, the finding that benefits of low-DOA varied as a function of multi-tasking are relatively robust as the interaction effects were of similar size across both this and Chapter 2 despite different samples and differences between the experimental paradigms. Robustness and replication have been noted as an important goal for transparency and verifiability of findings in psychology (Cumming, 2014; Pashler & Wagenmakers, 2012).

Replication of multi-tasking ability again predicted SA and subjective workload

SA results also were replicated from Chapter 2. Multi-tasking ability positively predicted SA, with similar effects sizes (Chapter 3 $\beta = -0.28$ versus Chapter 2 $\beta = -0.25$) despite increased query sensitivity and other differences between the present Chapter and Chapter 2. The replicated finding of multi-tasking predicting SA is also in line with other evidence for a link between multi-tasking and SA (Bender et al., 2017), which may be due to the working memory components shared by both (Redick et al., 2016). The robustness of the evidence that multi-tasking was related

to SA is provided in the consistent effect demonstrated in Chapters 2 and 3, as well as in the changes to sensitivity of SA queries in the current chapter which resulted in the same effect. Also consistent with Chapter 2, low-DOA did not affect SA, nor did the two variables interact.

It was expected that highlighting relevant aircraft with low-DOA to assist task performance would improve SA, as it would make task-relevant information more salient. However, neither this chapter nor Chapter 2 found low-DOA affected SA adding to previous mixed findings on automation use on SA (Kaber & Endsley, 2004). Low-DOA may simply not affect the aspects of the task environment assessed by SA queries. This explanation is supported by a meta-analysis that examined SA method sensitivity and found of the SPAM studies reviewed (n = 26) there was a sensitivity of 64%, but sensitivity was only 33% in the three ATC studies reviewed (Endsley, 2020). Comparatively, SAGAT (reviewed across 119 studies) showed an overall sensitivity of 89%, and this increased to 94% for ATC studies (n = 17). Although it is not clear from this review that SPAM is not sensitive specifically to manipulations involving automation, the overall comparison suggests that experiments which used SPAM may be less likely to find expected effects than those that used SAGAT, particularly in the ATC environment.

SA queries are aimed at inquiring about events in the ATC tasks and therefore awareness of such events should reasonably be related to task performance. However, low-DOA may have helped conflict detection without generally increasing awareness of events in the task. For instance, queries about the past or future location or one or more aircraft (e.g., 'next waypoint to be crossed…' or 'current sector location of…') may have aircraft highlighted if they are part of a conflict or near miss pair. But, at any given time there may be 2-3 other pairs of highlighted aircraft, making a specific highlighted aircraft hard to differentiate. Queries about the number of aircraft in a sector are not aided by low-DOA, as only aircraft in need of action (i.e., approaching a boundary or potentially in conflict) are highlighted; therefore, when answering these types of queries, some aircraft would be highlighted and others not. Queries assessing the number of aircraft that were (or will be) accepted/ hand-off in the last (or next) 30 seconds may have also not been aided by low-DOA as highlighting only occurs 20 seconds before action is required, and once accepted or handed-off, highlighting turns off. Finally, low-DOA does not highlight speed or flight level information as this is not relevant to acceptances or hand-offs. Therefore, queries regarding speed or flight level may have also not been aided. Such information is relevant to conflict detection, but it is not the criteria used to highlight potential conflict pairs. Thus, acquiring and maintaining awareness of the task may not have been assisted by low-DOA, even though queries assess information relevant to the tasks to which automation was applied.

Low-DOA reduced objective workload

Consistent with Chapter 2 both subjective workload measures showed that workload was rated lower in the low-DOA condition than in the manual condition. The addition of low-DOA to

conflict detection may explain the larger condition effect here (η^2 = 0.25) compared to in Chapter 2 (η^2 = 0.10). The pattern of results for subjective workload obtained with the single-item post-scenario measure were replicated with the more sensitive NASA-TLX measure, although the effect was smaller (η^2 = 0.15) possibly due to the missing data resulting in a smaller sample for this measure. The global score (i.e., workload averaged across subscales) showed lower ratings for the low-DOA compared to the manual condition. This effect of low-DOA was reflected in specific subscales for effort, mental demand, and temporal demand. These subscales relate to cognitive aspects of workload (as opposed to physical aspects) which supports the utility of automation in reducing the load on cognitive resources. The effect of low-DOA on subjective workload here mirrors the strong relationship found in previous literature between workload and task requirements (Matthews et al., 2020). A theoretical explanation of this effect posits that subjective ratings may be more sensitive to working memory requirements related to the task such that variations in task requirements are represented in working memory and thus accessible to participants when making relative judgements of workload (Vidulich & Tsang, 2012).

For subjective workload, the NASA-TLX was included in this chapter to better capture multi-tasking related considerations of workload. Underlying its inclusion was the expectation that better multi-taskers would rate their subjective workload lower because they have greater cognitive capacity (thus, also perform better – as shown above) and would find the task easier than poorer multi-taskers who have less mental capacity. This expectation was built on previous literature (see Vidulich & Tsang, 2012; Yeh & Wickens, 1988) that suggested that the determinants of subjective workload can be reduced to 1) task demands (e.g., difficulty, number of competing priorities), and 2) the attentional or processing resources a person brings to the task (e.g., perception, updating memory, decision making and response processing; Young & Stanton, 2001). Many of these information processing resources are required for multi-tasking (Redick et al., 2016) and are modulated by individual's abilities (Just et al., 2003).

From this understanding of determining factors, it is assumed that judgements of workload would be a combination of how much cognitive resources are used, and which cognitive resources are used. However, as noted by some (Ericsson & Simon, 1993) the second part of this assumption requires people to have insight into their information processing resources. Therefore, if participants lack insight into their cognitive abilities or do not relate those abilities to how they performed the task, their ratings of subjective workload would only constitute task demands. This appears to be the case from the present results as contrary to what was predicted here, but consistent with the findings in Chapter 2, responses to neither subjective workload measures differed as a function of multi-tasking ability. Thus, multi-tasking processes do not appear to reflect conscious considerations of participant's subjective ratings of workload.

In contrast to Chapter 2, objective workload was also not predicted by multi-tasking ability. This is not consistent with literature suggesting that objective workload may be more sensitive to variations in cognitive resources than subjective workload which set the expectation of finding a multi-tasking effect (Gawron, 2008; Matthews et al., 2020). It is unclear why multi-tasking did not impact objective workload here. However, given previous effects detected in Chapter 2 were small, a small effect may have been present here, but this Chapter would have needed greater statistical power to detect it. Further, objective workload was not affected by low-DOA. Workload has been inconsistently found to correlate with primary task performance as literature suggests objective workload may be related to primary task requirements only indirectly (Chapter 2; Hart & Wickens, 2010). Considering this it is not unexpected that low-DOA did not benefit objective workload.

Limitations

While this chapter contributes to our understanding of how complex tasks are benefited by low-DOA, provided key replication of previous findings in this thesis, and presented methodological improvements in SA and workload measures, there are limitations. The sample population of undergraduate students was chosen for practical reasons. It afforded a substantial sample that provided large statistical power to help develop the evidence for the theoretical ideas being developed. However, student novice populations differ from the expert populations that use automation in complex work environments such as ATC (Strybel et al., 2016). For instance, SA has been found to differ between air traffic controllers who rely heavily on automation to prevent cognitive overload and novices who try to build detailed but costly internal mental models (Chiappe et al., 2012). Experienced operators may self-select for occupations with high-level cognitive skills, which means ATC operators as a sub-population likely have a smaller range of multi-tasking abilities heavily skewed to the higher end of the cognitive continuum than the general population. It is therefore desirable for future studies to replicate these findings with expert operators to establish their generalisability.

Practical implications and future directions

The key finding here is that automation may not assist everyone equally. *Within* a single DOA there can be considerable variability in the benefits conferred by automation which depend on cognitive ability, not only *across* DOAs as has been traditionally studied. Low-DOA significantly benefits both simple (e.g., hand-off, acceptance) and complex (e.g., conflict detection) tasks within the ATC, suggesting it can be applied generally across a range of tasks representative of real-world automated environments. However, the magnitude of these benefits is also variable across tasks, as demonstrated by differences in the effect sizes of low-DOA found for conflict detection, acceptances, and hand-offs. Interactions with multi-tasking ability further

indicates that low-DOA may not confer the same benefits to all operators equally. The \ effect of interaction found for acceptances and hand-offs, and the low-DOA effect which extended to conflict detection in this chapter do not extend to SA and workload. This is unexpected as SA and workload have generally been found to relate to performance (Onnasch et al., 2014). The main driver of this dissociation may be the nature of the automation. Essentially, while low-DOA may have enough of an effect on performance to detect an interaction, the effects may not be strong enough to be detected for workload and SA, were as a higher-DOA may have stronger effects on all outcomes which may allow an interaction with multi-tasking ability to be detected.

This chapter had a more methodologically robust operationalisation of multi-tasking with a three-task factor underpinning the latent factor. It is interesting to note that despite this, the results reported here were very similar to Chapter 2 regarding multi-tasking and its interactions with low-DOA where these effects were detected previously. This suggests that both factor structures adequately captured multi-tasking and validates the use of the two-factor structure in Chapter 2. It has also extended the methodological validity of assessing multi-tasking as it predicted task performance, including interactions with automation, SA, and workload consistently as in Chapter 2.

The current chapter examined low-DOA which assisted multi-tasking in this task. Future studies could apply this methodology to the ATC and examine other DOAs to see if similar pattern of benefit occurs with medium or high-DOA. For example, if poor multi-taskers benefit a lot from low-DOA compared to manual, do they benefit even more from high-DOA? Is the benefit proportional to the DOA? The ATC represents the sort of tasks which are automated but may not generalise to all workplaces. Future studies could also examine low-DOA in a different task environment such as UAV, submarine, or nuclear power-plant simulators to generalise these findings to different task requirements.

A further future direction is to extend the similarity of the automation closer to real-world systems. The automation provided here is perfectly reliable, unlike real systems which are often highly reliable, but not infallible. It is possible that when low-DOA fails, poorer multi-taskers who received the greatest benefit from functioning automation may also experience commensurately greater costs to performance and workload when automation fails (e.g., either stops working or works incorrectly). Similarly, although low-DOA did not improve SA, low-DOA failures may lead to reduced SA as the distraction caused by intermittent automation failure may disrupt information being encoded in working memory or located efficiently in the environment. It is therefore important to understand what happens when automation fails and examine if the patterns reported in this chapter can be replicated, as automation failure is more reflective of real-world conditions.

Another future question is whether these findings generalise to cognitive abilities other than multi-tasking. Individual differences in other domains such as spatial awareness may show

different patterns of interaction with automation (Wright et al., 2018). Likewise, more specialized sub-component processes of multi-tasking such as working memory may interact differently with automation than higher-level processes (Jipp & Ackerman, 2016). Alternative cognitive abilities such as these can be readily profiled and may show alternative patterns of interaction with low-DOA which would be valuable for designers of future automated systems which will be tailored to individual's needs.

Conclusion

As cognitive ability increases the benefit of low-DOA to complex task performance is reduced, SA increases and objective workload declines. The present findings in conjunction with the previous chapter indicates low-DOA can confer benefits to operators with lower cognitive ability, at least in the domain of multi-tasking, across a range of simple and complex tasks in the ATC environment. Traditionally higher DOAs have been considered necessary to confer significant performance benefits. This chapter indicates this may not be the case if individual differences in cognitive ability are considered in the design of automated systems. Multi-tasking ability may be important to know about operators of automation, as underlying ability determines not only complex task performance, but workload and SA which are interrelated with safe operation of automation, particularly in aviation.

References

Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, O., Faraji, F., Janowich, J., Kong, E., Larraburo, Y., Rolle, C., Johnston, E., & Gazzaley, A. (2013). Video game training enhances cognitive control in older adults. *Nature*, 501(7465).

Bartholomew, D. J., Deary, I. J., & Lawn, M. (2009). The origin of factor scores: Spearman, Thomson and Bartlett. *British Journal of Mathematical and Statistical Psychology*, 62(3), 569-582.

Battiste, V., & Bortolussi, M. (1988, October). Transport pilot workload: A comparison of two subjective techniques. *In Proceedings of the Human Factors Society Annual Meeting*, 32(2), 150-154. Sage CA: Los Angeles, CA: SAGE Publications.

Bender, A. D., Filmer, H. L., Naughtin, C. K., & Dux, P. E. (2017). Dynamic, continuous multi-tasking training leads to task-specific improvements but does not transfer across action selection tasks. *Science of Learning*, 2(1), 1-10.

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment*, *19*(3), 354-369.

Bowden, V. K., Griffiths, N., Strickland, L., & Loft, S. (2021). Detecting a Single Automation Failure: The Impact of Expected (But Not Experienced) Automation Reliability. *Human Factors*.

Chiappe, D., Vu, K. P. L., & Strybel, T. (2012). Situation awareness in the NextGen air traffic management system. *International Journal of Human-Computer Interaction*, *28*(2), 140-151.

Chalder, T., Berelowitz, G., Pawlikowska, T., Watts, L., Wessely, S., Wright, D., & Wallace, E. P. (1993). Development of a fatigue scale. *Journal of psychosomatic research*, *37*(2), 147-153.

Chérif, L., Wood, V., Marois, A., Labonté, K., & Vachon, F. (2018). Multi-tasking in the military: Cognitive consequences and potential solutions. *Applied Cognitive Psychology*, *32*(4), 429–439.

Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, *19*(3), 259–282.

Chen, J. Y. C., & Terrence, P. I. (2009). Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment. *Ergonomics*, 52(8), 907-920.

Chen, S. I., Visser, T. A., Huf, S., & Loft, S. (2017). Optimising the balance between task automation and human manual control in simulated submarine track management. *Journal of Experimental Psychology: Applied*, *23*(3), 240.

Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance*, 21, 109–127.

Cullen, R. H., Dan, C. S., Rogers, W. A., & Fisk, A. D. (2014). The effects of experience and strategy on visual attention allocation in an automated multiple-task environment. *International Journal of Human-Computer Interaction*, *30*(7), 533-546.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, 25(1), 7-29.

Derry, K. L., Ohan, J. L., & Bayliss, D. M. (2017). Toward understanding and measuring grandiose and vulnerable narcissism within trait personality models. *European Journal of Psychological Assessment.*

Dikmen, M., & Burns, C. M. (2016, October). Autonomous driving in the real world: Experiences with tesla autopilot and summon. *In Proceedings of the 8th international conference on automotive user interfaces and interactive vehicular applications*, 225-228.

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Response time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin,* 145(5), 508.

Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., Nikolic, D., & Manning, C. A. (1998). Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly*, *6*(1), 1-20.

Durso, F. T., Dattel, A. R., Banbury, S., & Tremblay, S. (2004). SPAM: The real-time assessment of SA. *A cognitive approach to situation awareness: Theory and application*, *1*, 137-154.

Dux, P. E., Tombu, M. N., Harrison, S., Rogers, B. P., Tong, F., & Marois, R. (2009). Training Improves Multi-tasking Performance by Increasing the Speed of Information Processing in Human Prefrontal Cortex. *Neuron,* 63(1), 127-138.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. *In Proceedings of the Human Factors Society annual meeting*, 32(2) 97-101. Sage CA: Los Angeles, CA: Sage Publications.

Endsley, M. R. (2000a). Theoretical Underpinnings of Situation Awareness: A Critical Review Process.

Endsley, M. R. (2020b). The divergence of objective and subjective situation awareness: A meta-analysis. Journal of cognitive engineering and decision making, 14(1), 34-53.

Endsley, M. R. (2021). A Systematic Review and Meta-Analysis of Direct Objective Measures of Situation Awareness: A Comparison of SAGAT and SPAM. *Human Factors*.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, 37(2), 381-394.

Ericsson, K. A., & Simon, H. A, (1993), *Protocol Analysis: Verbal Reports as Data*, rev. ed., MIT Press, Cambridge, MA.

Falkland, E. C., & Wiggins, M. W. (2019). Cross-task cue utilisation and situational awareness in simulated air traffic control. *Applied ergonomics*, *74*, 24-30.

Fahlman, S. A., Mercer-Lynn, K. B., Flora, D. B., & Eastwood, J. D. (2013). Development and validation of the multidimensional state boredom scale. *Assessment*, *20*(1), 68-85.

Fothergill, S., Loft, S., & Neal, A. (2009). ATC-labAdvanced: An air traffic control simulator with realism and control. *Behavior Research Methods*, *41*(1), 118–127.

Galster, S. M., Duley, J. A., Masalonis, A. J., & Parasuraman, R. (2001). Air traffic controller performance and workload under mature free flight: Conflict detection and resolution of aircraft self-separation. International Journal of Aviation Psychology, 11(1), 71–93.

Gawron, V. J., (2008). *Human Performance, Workload, and Situational Awareness Measures Handbook*, 2nd ed., CRC Press, Boca Raton, FL.

Grier, R. A. (2015, September). How high is high? A meta-analysis of NASA-TLX global workload scores. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59(1), 1727-1731. Sage CA: Los Angeles, CA: SAGE Publications.

Gucciardi, D. F., Hanton, S., Gordon, S., Mallett, C. J., & Temby, P. (2015). The concept of mental toughness: Tests of dimensionality, nomological network, and traitness. *Journal of personality*, *83*(1), 26-44.

Gutzwiller, R. S., Wickens, C. D., & Clegg, B. A. (2019). The role of reward and effort over time in task switching. *Theoretical Issues in Ergonomics Science*, *20*(2), 196–214.

Hambrick, D. Z., Oswald, F. L., Darowski, E. S., Rench, T. A., & Brou, R. (2010). Predictors of multitasking performance in a synthetic work paradigm. *Applied cognitive psychology*, 24(8), 1149-1167.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *In Advances in psychology*, (52), 139-183. North-Holland.

Hart, S. G., & Wickens, C. D., (2010) Cognitive Workload. *In: NASA human integration design handbook* (HIDH). NASA, Washington, DC, 190–222

Heatherton, T. F., & Polivy, J. (1991). Development and validation of a scale for measuring state self-esteem. *Journal of Personality and Social psychology*, 60(6), 895.

Jipp, M., & Ackerman, P. L. (2016). The Impact of Higher Levels of Automation on Performance and Situation Awareness. *Journal of Cognitive Engineering and Decision Making*, *10*(2), 138–166.

Just, M. A., Carpenter, P. A., & Miyake, A. (2003). Neuroindices of cognitive Workload: Neuroimaging, pupillometric and event-related potential studies of brain work. *Theoretical Issues in Ergonomics Science*, 4(1-2), 56-88.

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*(2), 113-153.

Leiden, K.J., K opardekar, P. & Green, S., (2003). *Controller workload analysis methodology to predict Increases in airspace capacity.* Paper presented at the AIAA's 3rd Annual Aviation Technology, Integration, and Operations (ATIO) Tech, November 2003, Denver, Colorado (Reston, VA: AIAA).

Li, Y., & Burns, C. M. (2017). Modeling automation with cognitive work analysis to support human-automation coordination. *Journal of cognitive engineering and decision making*, 11(4), 299-322.

Lien, M. C., Schweickert, R., & Proctor, R. W. (2003). Task switching and response correspondence in the psychological refractory period paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(3), 692.

Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs (?). *Behavior Research Methods*, 51(1), 40-60.

Loft, S., Bolland, S., Humphreys, M. S., & Neal, A. (2009). A theory and model of conflict detection in air traffic control: Incorporating environmental constraints. *Journal of Experimental Psychology*: Applied, 15(2), 106–124.

Loft, S., Humphreys, M., & Neal, A. (2004). The influence of memory for prior instances on performance in a conflict detection task. *Journal of Experimental Psychology: Applied*, 10(3), 173.

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.

Luciani, D., Löwgren, J., & Lundberg, J. (2020). Designing fine-grained interactions for automation in air traffic control. *Cognition, Technology & Work*, *22*(4), 685-701.

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87.

Matthews, G., De Winter, J., & Hancock, P. A. (2020). What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. *Theoretical issues in ergonomics science*, *21*(4), 369-396.

Marschall, D. (1996). Effects of induced shame on subsequent empathy and altruistic behavior (Unpublished master's thesis). George Mason University, Fairfax, VA

Mogford, R. H., Guttman, J. A., Morrow, S. L., & Kopardekar, P. (1995). *The Complexity Construct in Air Traffic Control: A Review and Synthesis of the Literature. Missing some details??*

Moray, N. (1979). *Mental workload: Its theory and measurement*. New York: Plenum Press

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors*, 56, 476–488.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, *30*(3), 286-297.

Parasuraman, R., Cosenzo, K. A., & De Visser, E. (2009). Adaptive automation for human supervision of multiple uninhabited vehicles: Effects on change detection, situation awareness, and mental workload. *Military Psychology*, *21*(2), 270-297.

Parasuraman, R., Masalonis, A. J., & Hancock, P. A. (2000). Fuzzy signal detection theory: Basic postulates and formulas for analysing human and machine performance. *Human Factors*, 42, 636–659.

Pashler, H. (1984). Processing stages in overlapping tasks: Evidence for a central bottleneck. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(3), 358–377.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. Perspectives on psychological science, 7(6), 528-530.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink?. *Journal of experimental psychology: Human perception and performance*, 18(3), 849.

Redick, B. T. S., Shipstead, Z., Meier, M. E., Montroy, J. J., Hicks, K. L., Unsworth, N., … Hambrick, D. Z. (2016). Cognitive predictors of a common multi-tasking ability: Contributions from working memory, attention control, and fluid intelligence. *Journal of experimental psychology: General*, 145(11), 1473.

Remington, R. W., Johnston, J. C., Ruthruff, E., Gold, M., & Romera, M. (2000). Visual search in complex displays: Factors affecting conflict detection by air traffic controllers. *Human Factors*, 42, 349–366.

Rosenberg, M. (1965). Rosenberg self-esteem scale (RSE). *Acceptance and commitment therapy. Measures package*, 61(52), 18.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, 49(1), 76-87.

Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive Control of Cognitive Processes in Task Switching. *Journal of Experimental Psychology: Human Perception and Performance*, 27(4), 763–797.

Saqer, H., & Parasuraman, R. (2014). Individual performance markers and working memory predict supervisory control proficiency and effective use of adaptive automation. *International Journal of Human Factors and Ergonomics 55*, *3*(1), 15-31.

Schumacher, E. H., Seymour, T. L., Glass, J. M., Fencsik, D., Lauber, E. J., Kieras, D. E., & Meyer, D. E. (2001). *Virtually perfect time-sharing in dual-task performance: Uncorking the central cognitive bottleneck*. Psychological Science, 12, 101-108.

Spielberger, C. D., Sydeman, S. J., Owen, A. E., & Marsh, B. J. (1999). Measuring anxiety and anger with the State-Trait Anxiety Inventory (STAI) and the State-Trait Anger Expression Inventory (STAXI). Lawrence Erlbaum Associates Publishers.

Stankovic, S., Loft, S., Rantanen, E., & Ponomarenko, N. (2011). Individual differences in the effect of vertical separation on conflict detection in air traffic control. The International Journal of Aviation Psychology, 21, 325–342.

Strybel, T. Z., Chiappe, D., Vu, K. P. L., Miramontes, A., Battiste, H., & Battiste, V. (2016). A comparison of methods for assessing situation awareness in current day and future air traffic management operations: Graphics-based vs text-based online probe systems. *IFAC-PapersOnLine*, *49*(19), 31-35.

Thalmayer, A. G., Saucier, G., & Eigenhuis, A. (2011). Comparative validity of brief to medium-length Big Five and Big Six personality questionnaires. Psychological assessment, 23(4), 995.

Trapsilawati, F., Wickens, C., Chen, C. H., & Qu, X. (2017). Transparency and Conflict Resolution Automation Reliability in Air Trffic Control. In *19th International Symposium on Aviation Psychology*, 419.

Van Selst, M., Ruthruff, E., & Johnston, J. C. (1999). Can practice eliminate the psychological refractory period effect?. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(5), 1268.

Vidulich, M. A., & Tsang, P. S., (2012). "Mental Workload and Situation Awareness." *In Handbook of Human Factors and Ergonomics*, 4th ed., edited by G. Salvendy, 243–273. New York, NY: Wiley

Visser, T. A. W., Ohan, J. L., & Enns, J. T. (2015). Temporal cues derived from statistical patterns can overcome resource limitations in the attentional blink. *Attention, Perception, and Psychophysics*, *77*(5), 1585–1595.

Vuckovic, A., Sanderson, P., Neal, A., Gaukrodger, S., & Wong, B. W. (2013). Relative position vectors: an alternative approach to conflict detection in air traffic control. *Human Factors*, 55(5), 946-964.

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.

Wickens, C. D. 1984. *Engineering Psychology and Human Performance*. 1st ed. New York: Harper Collins.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: a synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, 50(3), 449-455.

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. B. (2010, September). Stages and levels of automation: An integrated meta-analysis. In *Proceedings of the Human Factors and ergonomics society annual meeting*, 54(4), 389-393. Sage CA: Los Angeles, CA: Sage Publications.

Wilson, M. S., Farrell, S., Visser, T. A. W., & Loft, S. (2018). Remembering to execute deferred tasks in simulated air traffic control: The impact of interruptions. *Journal of Experimental Psychology: Applied*, 24, 360-369.

Wilson, M. S., Strickland, L., Farrell, S., Visser, T.A.W., & Loft, S. (2020). Prospective memory performance in simulated air traffic control: Robust to interruptions but impaired by retention interval., *Human Factors*, 62, 1249-1264.

Wong, C. Y., & Seet, G. (2017). Workload, awareness and automation in multiple-robot supervision. *International Journal of Advanced Robotic Systems*, *14*(3), 1–16.

Wright, J. L., Chen, J. Y. C., & Barnes, M. J. (2018). Human-Automation Interaction for Multiple Robot Control: The Effect of Varying Automation Assistance and Individual Differences on Operator Performance. *Ergonomics*, 1–34.

Xu, X., & Rantanen, E. M. (2003). Conflict detection in air traffic control: A task analysis, a literature review, and a need for further research. In R. S. Jensen (Ed.), *Proceedings of the 12th International Symposium on Aviation Psychology*, (pp. 1289–1295), Wright State University Press.

Yeh, Y. Y., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors*, 30(1), 111-120.

Young, M. S., & Stanton, N. A. (2001). Mental workload: theory, measurement, and application, in W. Karwowski (Ed.), *International Encyclopedia of Ergonomics and Human Factors: Volume 1*, (pp 507 – 509). Taylor Francis.

Zhang, J., E, X., Du, F., Yang, J., & Loft, S. (2021). The Difficulty to Break a Relational Complexity Network Can Predict Air Traffic Controllers' Mental Workload and Performance in Conflict Resolution. *Human Factors*, *63*(2), 240-253.

CHAPTER 4

Down but not out: Evidence that failure of low-degree automation costs poor multi-taskers more, but partial reliability retains benefits

Chapter Abstract

Background: This chapter will examine the extent to which individual differences in multi-tasking ability modulate the cost of the failure of low-degree automation (DOA) to performance, workload, and situation awareness. Automation is rarely perfectly reliable. Failure of automation, such as missing critical events can have significant detrimental effects on performance. Previous chapters in this thesis have established that multi-tasking ability modulates the benefit of low-DOA in the Air Traffic Control (ATC) task. However, it is unknown whether the cost to performance when such automation fails also is modulated by cognitive ability. **Methods:** Multi-tasking ability was indexed using a latent factor from three cognitive tasks completed by one-hundred-and-thirteen undergraduate students. Participants completed two conditions of a simulated ATC task: once with no automation (manual) and once with partially reliable low-DOA which missed acceptance and hand-off events. Conflict detection automation was perfectly reliable. **Results:** Poorer multi-taskers performed worse than better multi-taskers on low-DOA error trials than in both the reliable low-DOA and manual trials. Multi-tasking ability modulated the cost to performance. Reliable low-DOA benefitted conflict detection performance, SA, and subjective workload. Higher multi-tasking ability led to better ATC performance, SA, and objective workload. Multi-tasking ability however did not interact with low-DOA for these outcomes. **Conclusions:** Participants with poorer multi-tasking ability suffered greater costs to performance when low-DOA failed. However, all participants retained benefit to performance, SA, and workload compared to manual despite the partial reliability of low-DOA compared to manual. Individual differences in multi-tasking ability robustly predicted performance, SA, and workload outcomes.

Introduction

Many workplaces require cognitively demanding work and interaction with technology which increasingly requires performing multiple tasks by strategically directing and dividing attention (Gutzwiller et al., 2019). When task demands are high, multi-tasking may overwhelm operator cognitive capacity, resulting in reduced performance and increased workload (Wickens et al., 2010). However, when faced with the same task demands, not all individuals experience the same performance degradations. Operators with more cognitive resources may experience less performance degradation than peers with fewer resources (Schumacher et al., 2001). This suggests that individual differences in cognitive abilities like multi-tasking could modulate outcomes such as task performance, operator workload, and situation awareness (SA) – defined by Endsley (1988) as the perception and comprehension of the environment and projection of its near-future states in complex environments.

A conventional approach to reducing the mental burden of multi-tasking in areas such as transportation (e.g., Dikmen & Burns, 2016), aviation (e.g., Mogford et al., 1995), and defence (Rovira et al., 2017), is to automate tasks previously performed partially or fully by humans (Parasuraman et al., 2000). If automation is reliable (i.e., performs accurately and as an operator expects (Onnasch, Ruff et al., 2014), then it may free-up cognitive resources (S. Chen et al., 2018; Kaber & Endsley, 2004). Automation can potentially improve routine performance (Endsley & Kaber, 1999; Manzey et al., 2012), and reduce subjective workload (Manzey et al., 2012; Rovira et al., 2007) compared to 'manual' performance without automated assistance.

Research suggests that performance, workload, and SA outcomes depend on the *degree of automation* (DOA: Wickens et al., 2010) employed. Low-DOA systems support human information acquisition or analysis, but the human retains control over decision-making and action implementation. In comparison, high-DOA systems assist decision-making and action implementation with little-to-no human input. Reliable low- and high-DOA can both benefit task performance and reduce workload relative to manual performance (Onnasch, Wickens et al., 2014). However, while low-DOA often benefits operator SA, high-DOA often reduces SA relative to manual performance (see Chapter 2 Introduction for further discussion).

While previous research has extensively examined automation benefits and costs *across* DOA (i.e., comparing groups provided with low or high-DOA; Jipp & Ackerman, 2016; Wright et al., 2018), to my knowledge, research has not yet examined how individual differences in cognitive ability, such as multi-tasking, impact outcomes *within* a single DOA. Previous chapters in this thesis probed this question by assessing a range of interrelated performance and operator state outcomes, including task accuracy and response times, situation awareness, and objective and subjective workload in a simulated Air Traffic Control (ATC) task in which participants were asked to accept aircraft entering their sector, hand-off aircraft leaving their sector, and prevent

aircraft violating separation standards (conflict detection). In separate conditions, acceptance and hand-off tasks were either assisted by low-DOA that highlighted relevant aircraft or were unassisted by automation. Investigations using this task focused on whether: a) low-DOA outcomes differed from manual outcomes, b) multi-tasking ability influenced outcomes, and c) whether these two effects interacted.

In both Chapters 2 and 3, multi-tasking ability was found to predict ATC task performance for acceptance and hand-off tasks and modulated the impact of low-DOA on acceptance and hand-off performance. These outcomes support the utility of profiling the cognitive abilities of automation operators in understanding their interaction with DOA. Multi-tasking inconsistently modulated conflict detection performance (Chapter 2 only) such that poorer multi-taskers performed worse. A low-DOA benefit to conflict detection was also observed (Chapter 3). Multi-tasking predicted SA, a finding which adds to the previous theoretical linking of SA to multi-tasking by their shared working memory requirements (Gutzwiller & Clegg, 2013; Redick, 2016). However, neither chapter found much evidence for a relationship between multi-tasking and workload.

Imperfect automation and cost to performance

While the previous Chapters demonstrated that multi-tasking ability modulated the benefit of reliable automation, a question that remains unanswered is how is human performance modulated by individual differences in cognitive ability when automation is imperfect? Perfectly reliable automation (i.e., automation that works as expected 100% of the time) would likely be rare in complex settings due to inherent limitations of complex technology and the unpredictable nature of dynamic work environments. Automation may be exposed to situations for which it was not designed, resulting in hardware failure (Onnasch, Wickens et al., 2014). Alternatively, automation may work as intended by the designer but not as the operator expected (Ockerman & Pritchett, 2002). Broadly, when automation fails, performance can decline and workload can increase relative to manual performance (i.e., costs; Van Acker et al., 2018). Experimental literature on automation error describes a 'miss' as an unintentional action of a system that does not alert the operator to relevant information, such as failing to detect an incoming aircraft (Cullen et al., 2014). Costs to performance can be considerable, causing decrements below manual levels. A previous study using an engine system gauge monitoring task where automation was at 50% reliability, both high and low-DOA showed performance decrements below manual, although high-DOA showed a greater reduction (Rovira et al., 2002a). However, while misses can negatively impact operators' response time to critical events (Wickens et al., 2005), imperfect automation may still benefits operators' overall performance, provided that automation reliability is greater than 70% (Wickens & Dixon, 2007).

Multi-tasking ability may influence two aspects of human performance with imperfect automation. First, multi-tasking ability may impact when or if operators detect automation errors (Strand et al., 2014), as more cognitively capable operators may rely less on automation and/or be less complacent in following its advice (Cak et al., 2020). Second, multi-tasking ability may impact how much capability the operator has in order to resume manual control after successfully detecting an automation error (Cullen et al., 2014). Those with better multi-tasking ability may detect automation errors sooner and have a greater ability to resume manual control and therefore be less vulnerable to the costs associated with imperfect automation. Conversely, those with poorer multi-tasking ability may take longer to detect errors and lack the skills to take back manual control following an automation error and thus perform more poorly with imperfect automation (see Körber et al., 2015).

The current study

Chapter 4 aims to extend our understanding of human-automation interaction by examining if multi-tasking ability modulates the impact of imperfect automation on complex task performance. This significantly expands the real-world validity of the previous chapter's findings by testing whether they extend to more realistic conditions involving imperfect automation. To this end, the acceptance, hand-off, and conflict detection tasks were automated as in Chapter 3, but with reduced automation reliability on the acceptance and hand-off tasks. Additionally, this chapter also examined whether multi-tasking ability modulated the impact of imperfect automation on SA and workload. Stress and trust in automation were also explored in the context of imperfect automation.

Participants completed one ATC condition manually and one with the assistance of imperfect low-DOA. As in previous Chapters, acceptances and hand-offs low-DOA consisted of increasing the salience of relevant aircraft by changing their colour and making them flash. However, in the current study, automation only highlighted 70% of aircraft entering or exiting the airspace. The remaining 30% did *not* change colour and flash when they approached or departed the control sector. Conflict detection automation remained perfectly reliable as in Chapter 3. This chapter can therefore examine the potential for changes to performance for a perfectly reliable task (i.e., conflict detection) when automation for another task is imperfectly reliable. As in previous chapters, adjusted RTs reflecting speed and accuracy were calculated for each performance measure. Measures of SA and workload were identical to Chapter 3, while multi-tasking ability was assessed using same methodology as Chapter 2.

In terms of the potential multi-tasking effects on ATC performance, SA, and workload the following predictions were made. In the manual and in the imperfect low-DOA condition when automation worked it was expected better multi-taskers would out-perform poorer multi-taskers, as was found under perfectly reliable conditions in Chapters 2 and 3. This would likely

also hold for the conflict detection task as found in Chapter 2, although the multi-tasking effect was only approaching significance in Chapter 3 which included perfectly reliable automation. Commensurate with the Chapter 2 and 3 findings, it was again expected that better multi-taskers would have greater SA, and lower subjective workload.

In terms of low-DOA effects, for acceptances and hand-offs performance was expected to be better in reliable trials low-DOA than in the automation error trials (30% of trials in the low-DOA condition). Performance in the automation error trials may be worse than manual performance. Given conflict detection showed a benefit of low-DOA when automated in Chapter 3, it was expected that low-DOA would again be beneficial in this experiment. If task performance is independent, conflict detection performance may be similar to Chapter 3, indicating that imperfect automation in one task does not have a detrimental effect on another task. Alternatively, if task performance is *interdependent* imperfect automation for acceptances may reduce performance on conflict detection performance compared to Chapter 3.

The effect of imperfect low-DOA on SA has not been previously examined, although imperfect high-DOA was found to result in poorer SA while driving (Strand et al., 2014). Imperfect automation may cost SA as the additional demand on attentional resources caused by intermittent automation errors may disrupt information from being encoded in working memory or located efficiently in the environment. Regarding workload, on the one hand, workload may be higher under low-DOA than manual due to the increasing cognitive load imposed by the need to monitor the imperfect automation of acceptances and hand-offs. On the other hand, the reliable low-DOA for conflict detection may reduce the overall workload in the automated condition compared to manual as conflict detection is the more effortful task. It is unknown whether overall workload would be better or worse under imperfect low-DOA compared to manual, given previous findings have been mixed (Chapters 2 and 3).

With respect to the key area of interest – the moderating impact of multi-tasking ability on automation – it was expected that multi-tasking would interact with reliability to modulate acceptance and hand-off performance, such that poorer multi-taskers would perform worse in the automation error trials than in the automation correct trials, thus demonstrating a greater cost of imperfect automation compared to better multi-taskers. If imperfect automation costs SA, it was also likely that low-DOA would be found to interact with multi-tasking ability such that poorer multi-taskers would have greater SA decrement when using low-DOA compared to manual, compared to better multi-taskers.

A final issue investigated was the impact of introducing imperfect automation on stress and trust and if this impact varied with individual differences in multi-tasking ability. Stress and trust have not been investigated in previous chapters but are introduced here as they are particularly relevant to investigate under conditions of automation error. As discussed in Chapter 1, imperfect automation may increase task-related stress, particularly in tasks which require

cognitive effort to monitor infrequent events (Warm et al., 2008). In turn, increased stress (Sauer et al., 2012) in automation is linked to poorer human performance using imperfect automation (Parasuraman et al., 2008). Stress may be due to task demands including performing multiple tasks at once (Sauer et al., 2012). For both these reasons, it was expected that low-DOA may interact with multi-tasking ability such that poorer multi-taskers experience greater levels of stress than better multi-taskers.

Inappropriate trust (see Bliss & Dunn, 2000) in automation is also linked to poorer human performance using imperfect automation (Parasuraman et al., 2008). Operators may become less trusting following automation errors (Wickens, Clegg et al., 2015). Trust has also been linked to individual differences in cognitive abilities. A study of attentional control in a military Command and Control simulator with miss and false alarm prone automation found those with less perceived attentional control ability had greater inappropriate trust in automation than those with higher ability (J. Chen, 2011). Those with poorer cognitive ability were therefore subject to worse performance outcomes when performing multiple tasks concurrently when automation failed (J. Chen, 2011). This is particularly relevant as participants were told in that the automation was imperfect in the present study. Therefore, it was expected that low-DOA may interact with multi-tasking ability such that poorer multi-taskers experience a greater reduction in trust than better multi-taskers.

Methods

Participants

University undergraduate students (N = 121) were recruited from a psychology research participation pool. Participants received AUD$10 and partial course credit. Eight participants' data were excluded: four due to missing data on one of the cognitive tasks or the ATC task, and four due to performing below chance accuracy on the Dual or PRP tasks. Thus, the final sample consisted of 113 participants (36 males, 75 females, two identified as 'other', M = 21.23, range = 18 – 47). The research was approved by the Human Research Ethics Committee of The University of Western Australia.

Measures

Multi-tasking: The multi-tasking battery used for this Chapter includes the same cognitive tasks used in Chapter 2.

Performance - Air Traffic Control Task (Fothergill et al., 2009): The task simulates ATC by presenting a dynamic flight sector through which aircraft travel along intersecting flight paths and at varying cruising altitudes and speeds. See Methods in Chapter 2 for more details on the ATC task.

Participants completed one low-DOA condition and one manual condition (counterbalanced). However, unlike the ATC task used in previous chapters, in this chapter low-DOA was imperfect, leading to automation errors (misses) for acceptances and hand-offs while remaining 100% reliable for conflict detection (see Figure 4.1, aircraft pair AA36 and QF94 are highlighted red for potential conflict). Of the 44-47 acceptance and 44-47 hand-off events (the numbers varied from person to person), aircraft in 14 of each type did not change colour and flash, yielding an overall automation reliability of ~70%. For example, as can be seen in Figure 4.1, aircraft EK30 is not yet accepted and EK18 is not yet handed-off; but both remain green illustrating an automation error. By comparison, aircraft NZ17 and NZ29 illustrate instances of reliable automation for conflict detection. Automation was reliable for all tasks for the first five minutes in each condition. After this, approximately every third aircraft ready for acceptance or hand-off was not highlighted by the automation. The distribution of automation errors was not predictable. However, there were constraints on the spacing of automation errors, as automation missed only one aircraft at a time entering or exiting the sector. See supplementary materials Table 9 for more details on timing of error events.

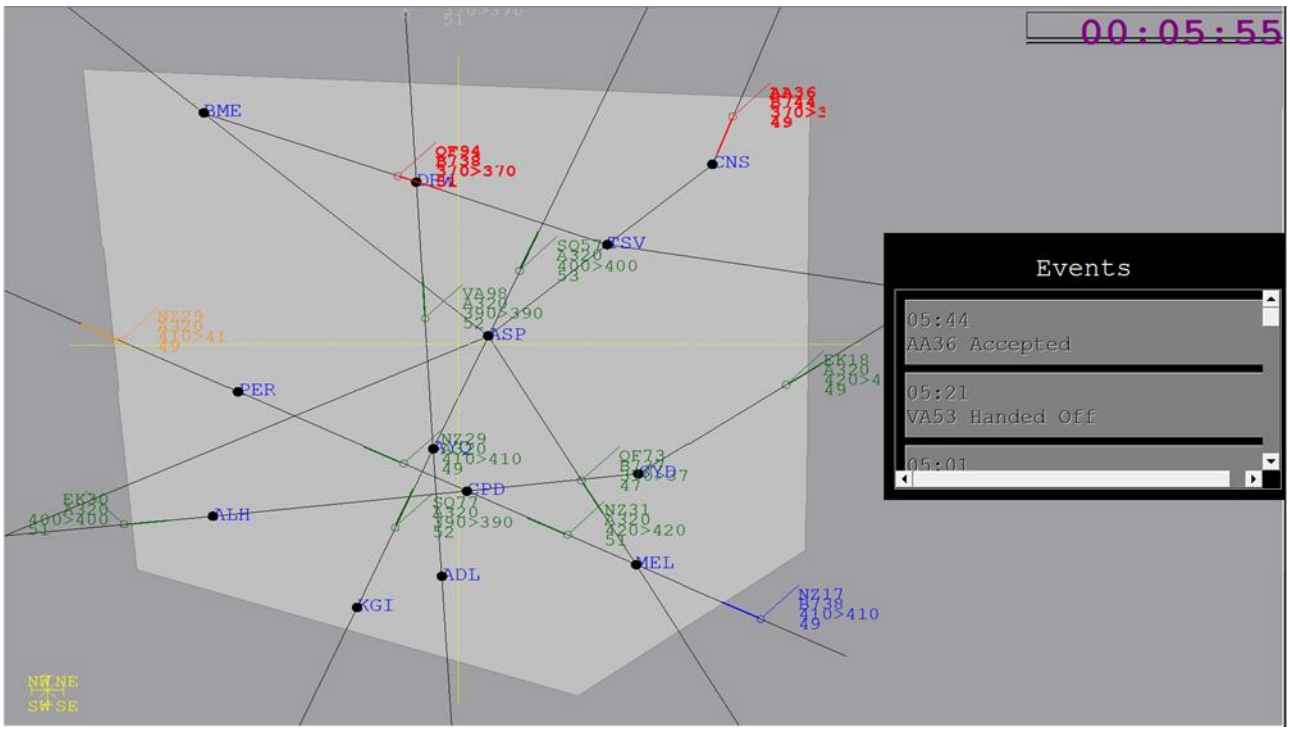


Figure 4.1. Air traffic control sector display. In the low-DOA imperfectly reliable condition (shown here), aircraft flashed blue when they approached the participants' sector (light grey polygon) indicating they required acceptance (i.e., NZ17) and flashed orange indicating they required hand-off (i.e., NZ29). Pairs of aircraft that may conflict are highlighted red (i.e., AA36 and QF94). In manual condition, all aircraft remained green. Actions performed by the participant were logged in the 'Events' box on the right of screen.

Situation Awareness: A modified Situation Present Awareness Method (SPAM; see Durso et al., 1998) was used to measure SA. The same queries were used as in Chapter 3.

Workload: Objective and subjective workload were assessed as Chapter 3.

Stress: The Short Stress-State Questionnaire (SSSQ; Helton, 2004) which is a modified version of the Dundee Stress State Questionnaire (Matthews et al., 1999) was completed before the first condition and after each of the two conditions. This measure comprises 24 items with responses on a 5-point Likert scale ('not at all' – 'a little bit' – 'somewhat' – 'very much' – 'extremely'). This is a well validated measure (Helton, 2004; also, Helton & Näswall, 2015) that examines engagement (Cronbach $\alpha = .81$), distress (Cronbach $\alpha = .87$), and worry (Cronbach $\alpha = .83$). Examples include "I expect to perform proficiently on this task' and 'I feel dissatisfied'.

Trust: Following each condition participants completed a questionnaire about their 'trust' in the automation (completed once for acceptances and hand-offs, once for conflict detection) adapted from the Trust in Automation scale (Merritt et al., 2013). These included six questions with responses on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). Examples include 'I believe the acceptance/hand-off/conflict automation is a competent performer'.

Other Questionnaires: Before the ATC task, participants completed a demographics questionnaire (including age, gender, video gaming experience, previous ATC experience) and a 9-item subscale from the Raven's Standard Progressive Matrices (RSPM; Bilker et al., 2012) to assess non-verbal intelligence.

Procedure

All cognitive tasks, ATC task and questionnaires were completed on a Windows PC with a BENQ 24$^{\text{-inch}}$ monitor running at a screen resolution of 1920 × 1080 and a refresh rate of 100Hz. Testing was conducted across two sessions. In the first session (60 minutes), participants were given information describing the experimental procedure and provided written consent. Participants first completed the demographics, then the cognitive task in counterbalanced order.

In the second session (120 minutes), participants completed the 20-minute training presentation outlining the three ATC tasks and the nature of the manual and automated conditions described in Chapter 2. Participants then completed a 20-minute practice condition under manual condition. Questionnaires were completed before the first condition and after each condition. Two 30-minute ATC conditions (manual and low-DOA) were completed in counterbalanced order. After the testing session, participants were debriefed, given the opportunity to ask additional questions, and received remuneration.

Results

Data cleaning

Before calculating a multi-tasking factor, the following data cleaning procedure was applied. For the PRP and Dual tasks and ATC task, mean target response times (RTs) were calculated only for trials with correct responses and RTs more than 150ms and less than 3 SD above the mean RT for each participant. Additionally, all participant data was omitted if they had an overall mean accuracy of less than 50%.

Dual task: Data cleaning resulted in 0.90% of Dual task data being removed.

PRP task: In addition to the data cleaning described above, PRP calculations omitted RTs if there was less than 50ms between keyboard responses (Tombu & Jolicœur, 2005). Mean RTs were calculated separately for each stimulus type (visual or auditory: PRP; Table 4.2). Data cleaning resulted in 13.2% of PRP trials removed.

AB task: Accuracy for T2 was calculated only for trials in which the first target (T1) was correctly identified. Mean T1 and T2|T1 accuracy were calculated separately as a function of lag (Raymond et al., 1992; see Table 4.3). No AB data was removed.

SA: mean RTs were calculated using correct response trials with RTs longer than 150ms or less than 3 SD above each participant's mean RT. This resulted in 1.33% of trials excluded. Adjusted RTs were calculated by dividing RT's by mean accuracy.

ATC task: Mean target response times (RTs) were calculated only for trials with correct responses and RTs more than 150ms and less than 3 SD above the mean RT for each participant. This resulted in the exclusion of 2.87% of acceptances, 3.31% of hand-off and 0% of conflict detection trials. As in previous chapters, adjusted RTs were calculated separately for acceptances, hand-offs, and conflict detection trials by dividing mean RTs by the corresponding mean accuracy. This combined accuracy and RT into a single composite measure accounting for speed-accuracy trade-offs (Liesefeld & Janczyk, 2019; Visser et al., 2015).

To account for the inclusion of 30% automation error trials, raw data for the acceptance and hand-off tasks was split by 'trial type' (reliable vs error trials) and processed separately using the outlier criteria described above. Each trial type was then adjusted by its own accuracy (i.e., 'error trials' adjusted by error trial accuracy which was calculated out of 14 or N minus number of outliers removed; likewise for reliable and manual trials). This did not result in the exclusion of any data. Then reliable and error trial sets with outliers removed were combined into an 'all trials' data set for the automated condition for each participant. Adjusted RTs were calculated by dividing the RT of each trial by the overall accuracy of 'all trials. This did not result in the exclusion of any data.

Workload: Objective workload was calculated as a mean RT for 'Ready for Question' prompts which were responded to before the 10-second timeout. Workload based on the NASA TLX scores were calculated in the prescribed way (Battiste & Bortolussi, 1988; Grier, 2015) with the ratings of each scale weighted by the number of times it was selected in the combinations (zero to five) to produce a weighted subscale score. A global workload score was calculated as the mean of the weighted subscale scores which was used as the dependent variable in analysis. The single-item workload measure had no data processing applied however two participants did not complete the measure leaving a sample of 111.

Table 4.1. Mean percentage of targets correctly identified and mean response time (in milliseconds) in the Dual Response Selection Task, separated by target type, with standard deviation in parentheses.

Trial Type	Shape		Sound	
	Accuracy	RT	Accuracy	RT
Single task	92.56 (6.76)	793 (117)	91.93 (6.50)	875 (145)
Dual task	89.06 (8.21)	988 (150)	98.06 (8.21)	1114 (145)

Table 4.2. Mean percentage of targets correctly identified and mean response time (in milliseconds) in the PRP task, separated by target type, with standard deviation in parentheses.

Inter-target interval	Sound trials		Visual trials	
	Accuracy	RT	Accuracy	RT
200ms	98.53 (1.77)	2093 (567)	98.42 (1.40)	1677 (490)
1000ms	97.17 (2.37)	1517 (479)	96.91 (3.71)	1738 (591)

Table 4.3. Mean percentage of targets correctly identified in the AB task with standard deviation in parentheses.

	T1	T2|T1
Lag 1	84.13 (11.16)	85.55 (11.21)
Lag 3	86.65 (11.00)	48.76 (22.20)
Lag 8	87.52 (10.00)	80.89 (15.13)

Table 4.4. Mean performance in the ATC task (acceptance, hand-off, conflict detection) and SPAM task as a function of condition. Low-DOA split into reliable and automation error trials for acceptance and hand-off. Accuracy is the percentage of correct trials. RTs are response times in seconds on correct trials. Standard deviations are in parentheses

	Manual		Low-DOA			
			Reliable Trials		Error Trials	
Task	Accuracy	RT	Accuracy	RT	Accuracy	RT
Acceptance	95.96 (10.13)	4.39 (1.74)	99.28 (9.53)	3.36 (1.60)	91.12 (13.63)	4.78 (2.10)
Hand-off	96.33 (10.35)	3.81 (1.68)	99.94 (0.00)	2.69 (1.53)	86.42 (10.70)	4.83 (2.08)
Conflict detection	93.57 (12.90)	113.16 (31.77)	98.39 (10.22)	90.08 (42.33)	-	-
SPAM	90.25 (12.28)	16.02 (4.99)	89.28 (16.37)	14.71 (3.95)	-	-

Latent factor analysis

For consistency with Chapter 3 the same three variables were used to formulate the latent factor; these included the mean RT at shortest inter-target interval in the PRP task (Short PRP RT), mean RT in the dual stimulus presentation condition in the Dual Task (Dual RT) and AB magnitude (described in Chapter 2). Individually, these tasks represent the more difficult condition within each task, thus reflecting multi-tasking ability. As previously an EFA was conducted using Principal Axis Factoring extraction and Direct Oblimin rotation. The loadings were acceptable (Dual = 0.74, uniqueness = 0.42; PRP = 0.48, uniqueness = 0.77; AB = 0.26, uniqueness = 0.93) although smaller than in the Chapter 3. The factor explained 28.7% of the variance. A multi-tasking factor score was created by for each individual by saving the Bartlett scores from the factor analysis which isolate shared variance on a factor across the tasks included (DiStefano et al., 2009). Scores were transformed (multiplied by -1) so that higher scores represented better multi-tasking ability (Bartholomew et al., 2009).

Linear mixed models

A series of linear mixed effects models (LMM) was conducted using the lme4 plugin (Bates et al., 2015) for R (RCore Team, 2015). Data was entered in a nested form to account for the within-subject design. The model predictors were chosen *a priori* to test the hypothesis of interest. Fixed factors entered were 'condition' with two levels (manual or low-DOA), multi-tasking score (continuous) and the interaction term of condition and multi-tasking. Random

effects entered were ‘participant number’ (intercept) and ‘condition by participant’ (slope) to control for within-subject variability across conditions.

To test for the effect of imperfect automation on acceptance and hand-off performance, trial type (manual, automation reliable and automation error) was included as a covariate, and an interaction term with trial type and multi-tasking was included. Model significance was assessed using a Chi-square test to compare to a null model which included only the random effects and the intercept (see supplementary materials Table 10). Unless otherwise stated, models significantly differed from the null. Assumptions of linearity, absence of autocorrelation, absence of influential cases and multicollinearity were satisfied (see supplementary materials Assumption plots). Homoscedasticity was at an acceptable level (i.e., evenly distributed) for all models. The assumption of normality of residuals was violated in all models, however LMM have been found to be robust to violations of this assumption (Knief & Forstmeier, 2021 see review by Schielzeth et al., 2020).

Models were also conducted with RSPM, previous experience with ATC, and counterbalance order as a covariate to control for potential effects of non-verbal intelligence, task order, or practice effects. Experience with the ATC task was a significant predictor of acceptance and hand-off only (see supplementary materials Table 11 and 12). None of the other covariates were significant. For brevity, effects of these covariates are not discussed further.

Standardised betas are presented in tables to allow comparison of effect sizes across studies and measures. Negative betas indicate a negative effect of multi-tasking meaning that adjusted RT decreased as multi-tasking scores increased. Likewise negative betas for condition indicate adjusted RTs decreased in the automated compared to the manual condition. Decreased adjusted RTs reflect better performance (i.e., more efficient).

ATC Performance: Accuracy and RT data for the ATC tasks and SPAM measure is presented in Table 4.4. Standardised coefficients and model fit indexes are presented in Table 4.5 (left column). Trial type significantly predicted acceptance and hand-off performance (see Figure 4.2) indicating a significant difference between manual, reliable automation trials and automation error trials for both tasks. Multi-tasking was a significant predictor only for hand-offs, indicating participants with higher multi-tasking scores handed-off aircraft better. Trial type significantly interacted with multi-tasking scores for both acceptances and hand-offs in the analysis which included manual, reliable and automation error trials. This result indicates that multi-tasking ability varied the effect of reliability such that poorer multi-taskers had larger differences in performance between their manual, reliable and automation error trials compared to better multi-taskers when all trial types are analysed together.

To investigate this interaction, separate LMM analyses were conducted for the acceptance and hand-off tasks that focused on comparisons of manual compared to reliable automation trials, and then manual compared to imperfect automation trials. Standardised coefficients and model fit

indexes are presented in Table 4.5 (middle and right columns). For both the acceptances and hand-off tasks, trial type was significant in all instances, indicating significant differences between manual and automation reliable trials, and between manual and automation error trials. As can be seen in Figure 4.2, this confirms that performance on manual trials was worse than on reliable automation trials, and that performance on manual trials was better than on imperfect automation trials. Additionally, as can be seen in Table 4.5, trial type interacted with multi-tasking ability for aircraft hand-off task, but not for acceptance. This further supports the finding that multi-tasking ability interacted with the effect of reliability such that poorer multi-taskers had larger differences in performance between their manual and reliable automation trials, as well as between their manual and automation error trials, compared to better multi-taskers for hand-offs. A greater cost of imperfect automation is therefore detected for poorer multi-taskers compared to better multi-taskers.

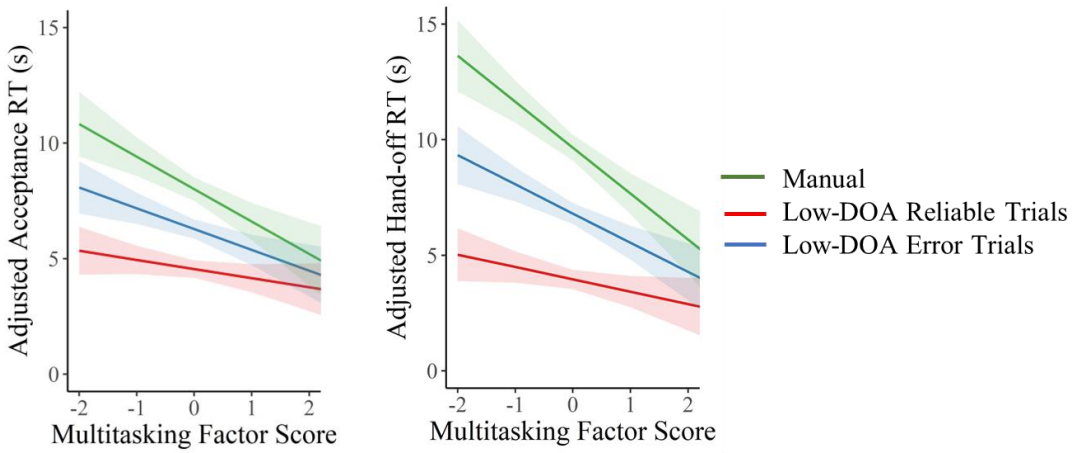


Figure 4.2. Acceptance and hand-offs by trial type – manual, low-DOA reliable and low-DOA error trials (measured in seconds) LMM slope and intercept. Black line represents manual condition, blue line represents low-DOA error trials and grey line represents low-DOA reliable trials. Shaded area represents 95% CI.

Table 4.5. Linear mixed effects standardised coefficient estimates, standard deviations in parantheses and model fit summaries.

Parameter	Acceptances Adjusted RT All	Acceptances Adjusted RT (Manual vs versus Reliable trials)	Acceptances Adjusted RT (Manual versus Error Trials)
Multi-tasking	-0.09 (0.24)	-0.08 (0.23)	-0.10 (0.25)
Trial Type	0.34*** (0.08)	-0.21*** (0.09)	0.09 *** (0.23)
Trial Type × Multi-tasking	-0.08*** (0.12)	0.02 (0.12)	-0.04 (0.30)
Observations	9629	8203	6193
Log Likelihood	-24782.92	-20301.85	-16583.50
AIC	49579.84	40617.70	33181.00
BIC	49630.05	40666.78	33228.10
R2 (conditional)	0.39	0.35	0.28
R2 (marginal)	0.13	0.05	0.02

Note: *** $p < .001$, ** $p < .01$, * $p < .05$. The estimates represent standardised coefficient estimates for fixed effects in the linear mixed model on the trial level specifying a random intercept and slope per participant. P-values were computed using the Wald approximation. R2 conditional is the model's total explanatory power. R2 marginal is the expanatory power of the fixed effects alone. Observations are the number of data points (trials) analyzed. Log likelihood is a measure of fit describing how well a parameter explains the observed data. AIC is Akaike Information Criterion, an extimator of prediction error. BIC is Bayesian Information Criterion, a measure for model comparison or selection.

Table 4.5. continued

Parameter	Hand-offs Adjusted RT All	Hand-offs Adjusted RT (Manual vs versus Reliable trials)	Hand-offs Adjusted RT (Manual versus Error Trials)
Multi-tasking	-0.11 (0.26)	-0.10 * (0.25)	-0.12 (0.27)
Trial Type	0.62 *** (0.27)	-0.22 *** (0.09)	0.19 *** (0.28)
Trial Type × Multi-tasking	-0.14 ** (0.35)	0.04 *** (0.11)	-0.05 * (0.36)
Observations	9644	8323	6060
Log Likelihood	-24748.20	-20069.98	-16561.95
AIC	49512.41	40153.96	33137.90
BIC	49569.81	40203.14	33184.86
R2 (conditional)	0.56	0.42	0.32
R2 (marginal)	0.32	0.05	0.05

Note: *** p < .001, ** p < .01, * p < .05. The estimates represent standardised coefficient estimates for fixed effects in the linear mixed model on the trial level specifying a random intercept and slope per participant. P-values were computed using the Wald approximation. R2 conditional is the model's total explanatory power. R2 marginal is the expanatory power of the fixed effects alone. Observations are the number of data points (trials) analyzed. Log likelihood is a measure of fit describing how well a parameter explains the observed data. AIC is Akaike Information Criterion, an extimator of prediction error. BIC is Bayesian Information Criterion, a measure for model comparison or selection.

For the conflict detection task, multi-tasking was a significant predictor, indicating participants with higher multi-tasking scores performed better (see Figure 4.3). Condition was also a significant predictor, indicating participants performed better in the low-DOA than the manual condition. However, the interaction between these factors was not significant. These results differ from Chapter 3 where multi-tasking was not found to predict conflict detection. However, condition significantly predicted performance in Chapter 3 and as was replicated here which suggests conflict detection was not interdependent on acceptances as was theorised. An alternative explanation is that the differences in automation reliability was detected by participants and participants placed appropriate trust in the reliable conflict detection automation (see discussions of trust below).

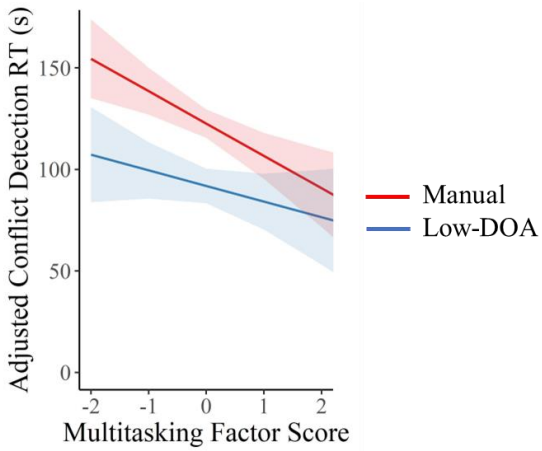


Figure 4.3. Conflict detection LMM slopes and intercepts. Performance measured via adjusted response time (seconds) against multi-tasking score (standardised), where a lower adjusted RT reflects better performance and higher multi-tasking scores reflect better multi-tasking ability. Shaded area represents 95% CI.

Table 4.6. Linear mixed effects standardised coefficient estimates, standard deviations in parantheses and model fit summaries.

Parameter	Conflict Detection Adjusted RT	SA Query Adjusted RT	Objective Workload
Condition	-0.21*** (3.65)	-0.07** (0.34)	0.008 (0.08)
Multi-tasking	-0.13*** (4.58)	-0.16*** (0.45)	-0.10* (0.11)
Condition × Multi-tasking	0.05 (4.69)	0.02 (0.43)	-0.0005 (0.10)
Observations	2180	3550	3482
Log Likelihood	-12236.00	-12314.70	-6980.00
AIC	24486.00	24642.40	13973.90
BIC	24525.80	24586.60	14017.00
R2 (conditional)	0.26	0.19	0.19
R2 (marginal)	0.06	0.03	0.00

Note: *** p < .001, ** p < .01, * p < .05. The estimates represent standardised coefficient estimates for fixed effects in the linear mixed model on the trial level specifying a random intercept and slope per participant. P-values were computed using the Wald approximation. R2 conditional is the model's total explanatory power. R2 marginal is the expanatory power of the fixed effects alone. Observations are the number of data points (trials) analyzed. Log likelihood is a measure of fit describing how well a parameter explains the observed data. AIC is Akaike Information Criterion, an extimator of prediction error. BIC is Bayesian Information Criterion, a measure for model comparison or selection.

Situation Awareness: The significant effect of multi-tasking (see Figure 4.4) indicated that participants with higher multi-tasking scores had better SA. However, there was no significant interaction with low-DOA. The multi-tasking finding, and lack of interaction replicates the findings of Chapter 2 and 3. The significant effect of condition indicated participants had higher SA in the low-DOA compared to the manual condition. This was the first time a condition effect was found across the three Chapters.

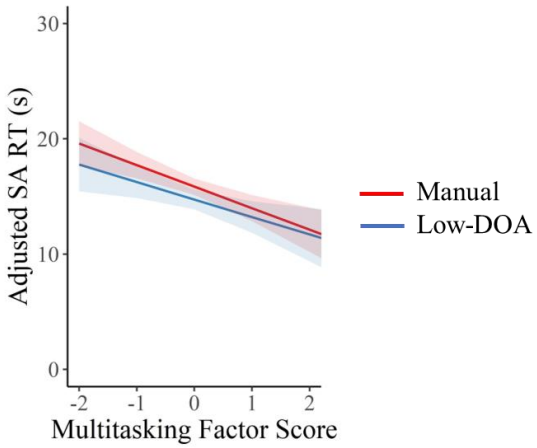


Figure 4.4. Situation Awareness LMM slope and intercept. Performance measured in adjusted response time (seconds) against multi-tasking score (normed). Black line represents manual condition and grey line represents low-DOA condition. Shaded area represents 95% CI.

Objective Workload: The model with multi-tasking did not significantly differ from the null. There was no significant main effect of condition or interaction.

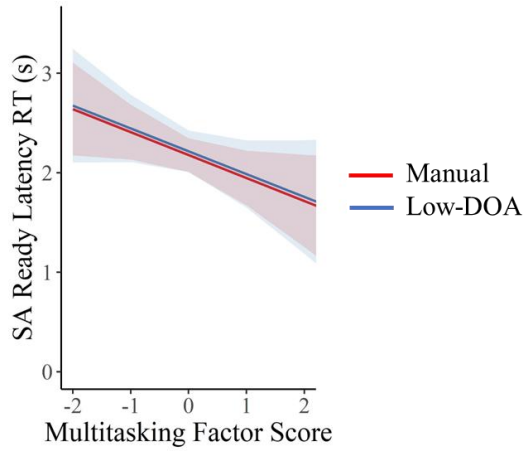


Figure 4.5. Objective workload measured by response latency to SPAM 'ready' prompt (measured in seconds) LMM slope and intercept. Black line represents manual condition and grey line represents low-DOA condition. Shaded area represents 95% CI.

Subjective Workload: As in previous Chapters, the single-item subjective workload measure was analysed using a repeated measures ANOVA with factors of condition (low-DOA and manual) and multi-tasking (low, medium, and high). As shown in Figure 4.6, this analysis yielded a significant effect of condition, $F(1, 108) = 48.81$, $p < .001$, $\eta^2 = 0.12$, with lower workload ratings in low-DOA condition compared to manual (automated: $M = 3.66$, SD = 1.34; manual: $M = 4.57$, SD = 1.11). There was no significant effects of multi-tasking ability or interaction ($F < 0.59$, $p > .55$, $\eta^2 < .007$). Subjective workload findings exactly replicate those obtained in Chapter 3.

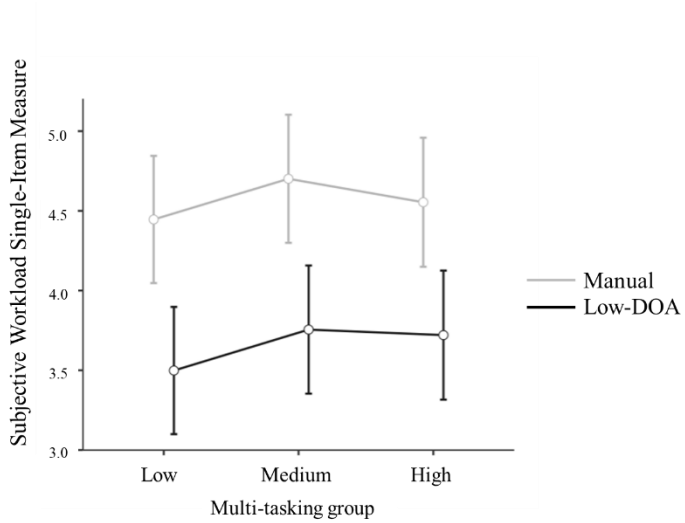


Figure 4.6. Subjective workload single-item measure across three groups of multi-tasking ability and two conditions of the ATC. Error bars represent 95% confidence intervals around the estimate of marginal means (between subjects).

The NASA-TLX total workload score was also analysed using a repeated measures ANOVA with condition and multi-tasking as factors. As seen in Figure 4.7, this yielded a significant effect of condition $F(1,110) = 46.31$, $p < .001$, $\eta^2 = 0.10$, with lower workload ratings in the automated compared to the manual condition. There were no significant effects of multi-tasking ability or an interaction ($F < 0.57$, $p > .39$, $\eta^2 < .01$). Identical analyses for each sub-component of the NASA-TLX yielded significant effects of condition for effort $F(1,110) = 31.87$, $p < .001$, $\eta^2 = 0.05$; mental demand $F(1,110) = 73.89$, $p < .001$, $\eta^2 = 0.16$; and temporal demand $F(1,110) = 10.03$, $p < .05$, $\eta^2 = 0.02$, each indicating higher workload in the manual condition (see Table 4.7 for mean and SD by condition). For each of the other sub-components there was no significant effect of condition ($F < 0.68$, $p > .41$, $\eta^2 < .001$). There was no multi-tasking effect or interaction for any sub-components ($F < 2.25$, $p > .11$, $\eta^2 < .001$).

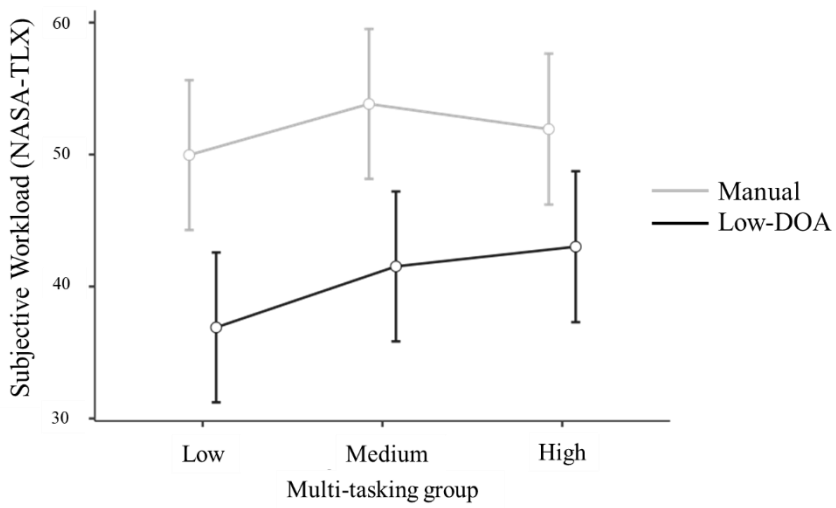


Figure 4.7. Subjective workload measure (NASA-TLX) across three groups of multi-tasking ability and two conditions of the ATC. Error bars represent 95% confidence intervals around the estimate of marginal means (between subjects).

Table 4.7. NASA-TLX Component mean and standard deviations across manual and low-DOA conditions

	Manual		Low-DOA	
	M	SD	M	SD
Effort	186	107	135	99.3
Mental Demand	248	120	148	104
Temporal Demand	135	103	102	94.4
Physical Demand	8.89	35.9	8.10	29.3
Performance	165	120	174	132
Frustration	36.4	74.4	39.6	88.2
Workload Total	51.9	18.1	40.5	17.2

Stress and trust: To investigate the effect of imperfect automation on participant's stress, the SSSQ was analysed in the same way as subjective workload. A repeated measures ANOVA with condition of 2 levels (low-DOA and manual) and multi-tasking treated as three approximately equal groups was conducted on stress score after each condition (averaged across items), with pre- task stress included as a covariate. As seen in Figure 4.8, this yielded a significant interaction between multi-tasking group and condition $F(2,108) = 3.18$, $p < .05$, $\eta^2 = 0.003$, with lower stress in automated compared to the manual condition for the low multi-tasking group. There was no significant effects of multi-tasking ability or condition ($F < 0.30$, $p > .07$, $\eta^2 < .01$). A follow-up paired sample t-test confirmed stress was lower in the automated condition ($M = 2.37$) compared to the manual condition ($M = 2.49$) for the low multi-tasking group ($t(36) = 3.55$, $p < .001$, Cohen's d = .58). Identical paired sample t-tests for the medium and high groups were non-significant ($t < -0.08$, $p > .50$, Cohen's d < .10).

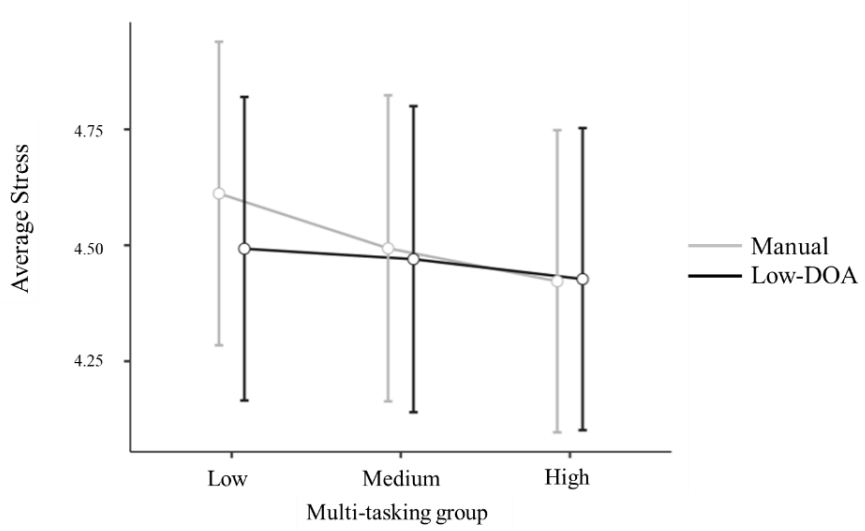


Figure 4.8. Post condition stress ratings across three groups of multi-tasking ability and two conditions of the ATC. Error bars represent 95% confidence intervals around the estimate of marginal means (between subjects).

Trust in automation was examined using a repeated measures ANOVA with task of 2 levels (trust in acceptance/hand-off automation and trust in conflict detection automation) and multi-tasking treated as three approximately equal groups. This yielded a significant effect of task, $F(1, 108) = 15.34$, $p < .001$, $\eta^2 = 0.048$, indicating participants rated the reliable conflict detection automation more trustworthy ($M = 3.19$, $SD = 1.09$) than the imperfect acceptance and hand-off automation ($M = 2.70$, $SD = 0.98$). As seen in Figure 4.9, a significant interaction between multi-tasking group and task was also detected, $F(1, 108) = 5.83$, $p < .05$, $\eta^2 = 0.037$, with higher multi-tasking ability participants rating the conflict detection automation more trustworthy than poor and medium multi-taskers. A follow-up paired sample t-test confirmed trust was lower for acceptance/hand-offs ($M = 2.47$) compared to the conflict detection ($M = 3.50$) for the high multi-tasking group (t (36) = -4.87, $p < .001$, Cohen's d = .80). Paired sample t-tests for the low and medium groups were non-significant, ($t < -0.80$, $p > .35$, Cohen's d < .14). This indicates higher multi-taskers were able to discern the difference in automation reliability better than other multi-taskers who did not differentiate their trust in either the reliable or imperfect automation. The effect of multi-tasking group was non-significant ($F = 0.49$, $p = .61$).

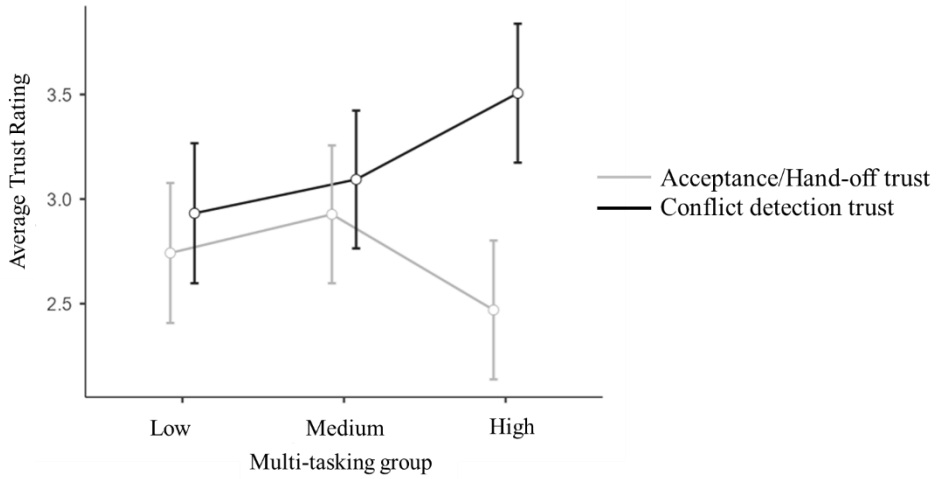


Figure 4.9. Post-automated condition average trust ratings across groups levels of multi-tasking ability and two levels of reliability for the tasks in the ATC. Error bars represent 95% confidence intervals around the estimate of marginal means (between subjects).

Discussion

This chapter aimed to examine how imperfectly reliable automation interacts with individual differences in multi-tasking to predict outcomes including task performance, workload, and SA in a simulated ATC environment. Previous chapters have established that multi-tasking ability modulated the benefits of perfectly reliable automation to performance such that automation conferred the most benefit to poorer multi-taskers compared to their manual performance. The present chapter builds on this foundation by assessing task outcomes using the same paradigm with imperfect low-DOA (i.e., automation 'miss' error on 30% of trials) to examine if multi-tasking ability also modulated the impact of automation errors on performance, SA, and workload when reliability was less than perfect. This chapter also examined whether use of reliable automation on a separate task (conflict detection) was influenced by use of imperfect automation on concurrent tasks, and the joint impact of imperfect automation and multi-tasking ability on stress and trust.

Participant's multi-tasking ability interacted with automation, such that acceptance and hand-off performance for automation error trials was worse than performance on both reliable automation trials and manual trials, and these effects were greater for participants with poorer multi-tasking ability. Multi-tasking ability also predicted better conflict detection performance, SA, and lower objective workload in line with Chapters 2 and 3, demonstrating a consistent effect of individual differences in cognition for these outcomes (although no interaction with automation). Conflict detection performance was greater, SA was better, and subjective workload was lower with low-DOA compared to manual performance. However, multi-tasking ability did not interact with condition despite the presence of both main effects for conflict detection and SA.

Multi-tasking ability varies the cost of imperfect automation

These findings support expectations that the cost of imperfect automation would be greater for poorer multi-taskers. Overall, a reduction in automation reliability of 30% had the most significant impact on operators who were least capable of overcoming such failure. The cost was largest for poorer multi-taskers when comparing reliable automation trials to manual, and manual to error trials for the hand-off task. While many studies have found imperfect automation has a greater detrimental effect on performance when using decision automation (high-DOA) than information automation (low-DOA: see McGarry et al., 2003, Rovira et al., 2002a; 2002b), few previous studies have examined automation failure with low-DOA compared to a manual condition. Di Nocera et al., (2005) found participants were slower to detect system errors in a simulated space operations task when they switched between manual and imperfect low-DOA, compared to when they used only imperfect automation and were not exposed to manual. The authors suggested there was a greater cost to changing between conditions because participants set their level of trust and reliance based on learned experience with the level of reliability. The present findings support this assertion, while also showing individual differences in cognitive ability can modulate performance and trust (see below).

This chapter found performance with imperfect automation was reduced below the level of manual performance for all participants. Two early studies by Rovria et al (2002a, 2002b) found a similar pattern in which accuracy and RT performance with imperfect automation (< 60%) was reduced below manual., This was found in both an engine system monitoring task, in which the imperfect automated task involved detecting malfunctions in gauges, while also performing concurrent tracking and fuel management tasks, as well as in a sensor-to-shooter Command and Control task (Rovira et al., 2002b). However, in both studies the cost to performance was greater for higher DOA than lower DOA. Thus, the present findings add to the literature indicating a cost to imperfect automation with low-DOA which may impact poorer multi-taskers more than good multi-taskers such that they perform worse than with no automation.

This chapter also found a cost to subjective trust in automation such that poorer multi-taskers rated the imperfectly reliable acceptance and hand-off automation with the same level of trust as the reliable conflict detection automation, while good multi-taskers showed less trust in the imperfectly reliable automation. This suggests poorer multi-taskers may have been insensitive to the difference in automation reliability between the acceptance/hand-off and conflict detection tasks. This could partly explain why better multi-taskers performed better on the conflict detection task: they recognised the automation for that task was more reliable than for acceptances and hand-offs and placed appropriate trust in that automation (Moray et al., 2000). In turn, this may have freed up cognitive resources which were used to monitor the acceptance and hand-off tasks to detect automation errors. By contrast, poorer multi-taskers who may not have recognised conflict detection automation was more reliable than acceptance and hand-off automation may

have wasted cognitive resources verifying the information of both tasks' automation (or none of it) which reduced their performance on both tasks compared to better multi-taskers.

Benefits to performance, SA and workload persist despite imperfect reliability

Studies which have examined the impact of imperfect low-DOA compared to manual suggest participants continued to use and derive a benefit to performance from imperfectly reliable automation (Cullen et al., 2013; Wickens & Dixon, 2007). Interestingly, the current chapter also found a benefit of imperfectly reliable automation on SA compared to manual. Thus, like performance, SA also appears to be able to benefit from imperfect low-DOA. At first glance, this seems unusual given that SA was not impacted by perfectly reliable automation in previous chapters. A potential explanation is that reliable low-DOA in the previous experiments may have discouraged participants from focusing on much of the display information, thus reducing their SA. With imperfect low-DOA, participants likely needed to pay greater attention to the display information as part of monitoring aircraft requiring acceptance or hand-off, thus increasing the speed of their responses to SA queries.

As in previous chapters, participants rated their subjective workload as lower under low-DOA than manual conditions, despite the imperfect automation reliability. Thus, the benefit of imperfect automation to workload parallels that found for performance as participants not only performed the tasks in the ATC better with imperfect low-DOA compared to manual, but they also reported a subjective sense that it was easier. The reliable conflict detection automation may also explain this, as this would be expected to reduce workload substantially given conflict detection is the more challenging task. This is a similar finding to previous aviation and ATC task studies which found imperfectly reliable information automation resulted in lower subjective workload than no automation (Karpinsky et al., 2018) and that subjective workload was not affected by automation errors (Di Norcera et al., 2006).

Objective workload did not show any effects of automation or multi-tasking ability. This differs from the previous chapters where significant multi-tasking effects were detected and differs from the findings for subjective workload which only detected significant condition effects. With respect to the divergence between objective and subjective workload measures Matthews et al. (2018) note that often different workload measures fail to converge. The authors consider this symptomatic of either a problem with workload as a unitary concept which is supported by a critical analysis of workload literature by Van Acker et al., (2018), or that different workload measures may index constructs other than workload, such as self-regulation and metacognition for subjective workload measures in performance settings. Here, it may be the case that different factors influenced subjective and objective workload ratings. As suggested in Chapter 3, the lack of a relationship between multi-tasking ability and subjective workload may be because multi-tasking ability is not something participants had much meta-awareness about, or

was, in fact, not considered relevant for participants when making their subjective ratings. By comparison, objective workload might theoretically be expected to relate to multi-tasking more directly as SPAM latency can be considered an index of spare cognitive capacity (Strybel et al., 2016), particularly with imperfect automation which may use more mental resources than perfectly reliable automation. However, this was also not the case presently, nor was it found in previous chapters with perfectly reliable automation. Thus, a divergence can be seen between workload measures, and between what was expected to be found for those measures based on theoretical literature (see Chapter 5 for further discussion).

Finally, there was a slight benefit to the subjective stress experienced by poorer multi-taskers who rated their stress as lower in the automated condition compared to the manual condition. All other multi-taskers rated their stress the same in both conditions. This suggests poorer multi-taskers felt less stressed when assisted by automation, even when for one task it was imperfectly reliable, supporting the cognitive benefits discussed above. Reduced stress for the poorer multi-taskers is also consistent with the lower subjective workload ratings for the automated condition, although this was not found to differ with multi-tasking ability. These stress findings differ from other research which suggests automation (including low-DOA) can increase stress in long tasks (30 min or more) by reducing motivation, concentration, and energetic arousal resulting in fatigue compared to manually performing a task (McGarry et al., 2003).

Limitations

The current chapter successfully showed the benefits and costs of imperfect automation can vary with individual differences in multi-tasking ability in a sample of undergraduate students who were novices to the ATC task. This sample was chosen for convenience and because it was possible to collect a large sample to maximize statistical power to find effects. However, this sample does not represent typical automation operators in the workplace who often have significant experience and skill honed over many years of interacting with a complex system which results in greater performance than novices (Balfe et al., 2015; Jamieson & Skraaning, 2018). Expertise has also been shown to determine outcomes when automation fails, as experts can recover performance more efficiently (Roth et al., 2019). In addition, experts may have different perceptions of workload (Matthews et al., 2020) and may be more trusting of automation than novices because they have more capacity to regain control of a system if it fails (Niu et al., 2018). Thus, the present pattern of findings could differ if experts rather than novices were examined.

Another limitation is the low-DOA deployed in this and the previous chapters. As previous literature has shown (Jipp, 2016; McGarry et al., 2003; Rovira et al., 2002a) higher DOAs are associated with greater costs when automation is imperfect than is found for lower DOAs (also see Wickens & Dixon, 2007). Higher DOAs provide more assistance, and greater

performance benefits when reliable (Kaber, 2018; Onnasch, Ruff et al., 2014; Wickens et al., 2010), but can cause overreliance and commensurately worse outcomes when automation fails (for review see Onnasch, Ruff et al., 2014). The present findings with low-DOA therefore may not generalise to high-DOA, particularly for imperfect automation. As complex systems more often employ higher DOA so that operators can cope with increased cognitive load (e.g., Chérif et al., 2018 in military context and Trapsilawati et al., 2017 in ATC), this could limit the application of these findings to real world systems.

Practical implications and future directions

This chapter examined the cost of imperfectly reliable low-DOA and how it differentially affected operators based on their multi-tasking ability. This approach was designed to maintain ecological validity by simulating the real-world systems which often work correctly but do not account for every eventuality (hence automation 'misses' critical events). Of course, the present findings are based on a limited type of error with a low-DOA, and thus may not generalise to automation failure situations such as the breakdown of systems due to technical failure or sabotage. Future experiments could vary the nature of the automation failure such as a full system breakdown in which automation stops working completely which may increase the likelihood of less capable operators experiencing greater cost to performance. This would help to establish the boundary conditions of how well people with varying levels of multi-tasking ability would compensate for the loss of automation, and how well they can return to performing a task manually.

Another question that would improve understanding of real-world automation errors would be to examine the cost of automation false alarms (i.e., inappropriate, or unnecessary alerts) and how they may interact with participants capabilities compared to misses. Previous literature has suggested false alarms have differential impacts on performance compared to miss events due to their salience (J. Chen & Terrence, 2009; Wickens et al., 2010). In the ATC environment this could involve automation of conflict detections which 'should' only highlight actual conflicting pairs of aircraft, but intermittently also incorrectly highlights near-miss pairs. It is possible that automation costs caused by false alarms may have a commensurately larger impact on operators with poorer cognitive ability than 'misses' found in the current chapter as they may be slower to identify the more plausible false alarm situation, which involves adjacent aircraft that must be inspected carefully, than the more obvious miss. In sum, a more nuanced approach to understanding the impact of automation failure is needed as failures can take multiple forms and these may interact differently with individual differences in cognitive abilities.

The present ATC task is relatively slow paced and includes a narrow set of repetitive tasks and low-DOA was limited to the acceptance and hand-off tasks which required target detection and a simple keyboard response. While these task conditions and automation

implementation accurately map onto many real-world tasks, it would nonetheless be useful to examine tasks that are fast-paced and dynamic, that require quick decision making and responding under uncertain conditions, like driving. It is unknown whether the cost to performance for less capable operators would be greater if the task was difficult and ambiguous. However, tasks with more dynamic and rapidly changing conditions and/or unpredictable elements may put greater stress on operators' multi-tasking ability, particularly the requirement to execute rapid task and attentional switches based on changing operational priorities. This might be expected to increase the magnitude of interactions between automation and individual differences in cognition, and perhaps reveal similar interactions in estimates of workload and SA.

Finally, future studies could further explore how participants trust in imperfect automation impacts performance, and whether this is modulated by cognitive ability. As noted earlier, this chapter's results suggest that poor multi-taskers may perform worse in part because they are less able to recognize automation failures and thus develop appropriate levels of trust. To examine this idea further, studies could assess multiple dimensions of trust, as described by Marsh and Dibben's (2003) three level model of dispositional (i.e., trait), situational and learned trust which could be examined by having participants identify how reliable the automation was as a percentage. This would allow for more precise examination of the accuracy of participant's judgement of reliability, and how this relates to dispositional trust, use of imperfect automation, and individual differences in cognitive ability.

Conclusions

Individual differences in cognitive ability have been shown to vary the benefit of low-DOA, and the cost to performance when automation is imperfectly reliable. This research showed 30% of missed events has a differential cost impact depending on participant's multi-tasking ability, which for the poorest multi-taskers resulted in performance worse than without automation. When considering low-DOA's overall effect, a consistent benefit was found to all ATC tasks, SA, and subjective workload, despite the imperfect reliability. This strengthens the rationale for considering individual's cognitive abilities when determining what automation should be deployed in the workplace, as those who gained a benefit from reliable automation, as shown in previous experiments and replicated here, also suffered the greatest cost when automation made an error.

References

Balfe, N., Sharples, S., & Wilson, J. R. (2015). Impact of automation: Measurement of performance, workload and behaviour in a complex control environment. *Applied Ergonomics*, 47, 52–64.

Bartholomew, D. J., Deary, I. J., & Lawn, M. (2009). The origin of factor scores: Spearman, Thomson and Bartlett. *British Journal of Mathematical and Statistical Psychology*, 62(3), 569-582.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv* preprint arXiv:1406.5823.

Battiste, V., & Bortolussi, M. R. (1988). *Assessment of pilot workload with the introduction of an airborne threat-alert system* (No. 881385). SAE Technical Paper.

Bender, A., Loft, S., Lipp, & Visser, T.A.W. (2018). *Advancing our understanding of warfighter cognition: Development of a "cognitive profiling" tool to enhance situation awareness.* Defence Science and Technology (DST) Group Human Performance Research Network (HPRnet).

Bilker, W. B., Hansen, J. A., Brensinger, C. M., Richard, J., Gur, R. E., & Gur, R. C. (2012). Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment*, *19*(3), 354-369.

Bliss, J. P., & Dunn, M. C. (2000). Behavioural implications of alarm mistrust as a function of task workload. Ergonomics, 43, 1283–1300.

Cak, S., Say, B., & Misirlisoy, M. (2020). Effects of working memory, attention, and expertise on pilots' situation awareness. Cognition, Technology and Work, 22(1), 85–94.

Chen, J. Y. (2011). Individual differences in human-robot interaction in a military multitasking environment. *Journal of Cognitive Engineering and Decision Making*, 5(1), 83-105.

Chen, J. Y. C., & Terrence, P. I. (2009). Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multi-tasking environment. *Ergonomics*, *52*(8), 907–920.

Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, 19(3), 259–282.

Chérif, L., Wood, V., Marois, A., Labonté, K., & Vachon, F. (2018). Multi-tasking in the military: Cognitive consequences and potential solutions. *Applied Cognitive Psychology*, *32*(4), 429–439.

Cullen, R. H., Dan, C. S., Rogers, W. A., & Fisk, A. D. (2014). The effects of experience and strategy on visual attention allocation in an automated multiple-task environment. International *Journal of Human-Computer Interaction*, 30(7), 533-546.

Di Nocera, F.., Fabrizi, R., Terenzi, M., & Ferlazzo, F. (2006). Procedural errors in air traffic control: effects of traffic density, expertise, and automation. *Aviation, space, and environmental medicine*, 77(6), 639-643.

Di Nocera, F., Lorenz, B., & Parasuraman, R. (2005). Consequences of shifting from one level of automation to another: main effects and their stability. *Human Factors in design, safety, and management*, 363-376.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation*, *14*(1), 20.

Dikmen, M., & Burns, C. (2016). Autonomous Driving in the Real World: Experiences with Tesla Autopilot and Summon. *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (AutomotiveUI '16), Ann Arbor, MI, USA., May, 225–228.

Durso, F. T., Dattel, A. R., Banbury, S., & Tremblay, S. (2004). SPAM: The real-time assessment of SA. *A cognitive approach to situation awareness: Theory and application*, *1*, 137-154.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society annual meeting*, 32(2), 97-101. Sage CA: Los Angeles, CA: Sage Publications.

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, 42(3), 462-492.

Fothergill, S., Loft, S., & Neal, A. (2009). ATC-labAdvanced: An air traffic control simulator with realism and control. *Behavior Research Methods*, *41*(1), 118–127.

Grier, R. A. (2015, September). How high is high? A meta-analysis of NASA-TLX global workload scores. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59(1), 1727-1731. Sage CA: Los Angeles, CA: SAGE Publications.

Gucciardi, D. F., Hanton, S., Gordon, S., Mallett, C. J., & Temby, P. (2015). The concept of mental toughness: Tests of dimensionality, nomological network, and traitness. *Journal of personality*, *83*(1), 26-44.

Gutzwiller, R. S., & Clegg, B. A. (2013). The role of working memory in levels of situation awareness. Journal of Cognitive Engineering and Decision Making, 7(2), 141-154.

Gutzwiller, R. S., Wickens, C. D., & Clegg, B. A. (2019). The role of reward and effort over time in task switching. *Theoretical issues in ergonomics science*, *20*(2), 196-214.

Helton, W. S. (2004, September). Validation of a short stress state questionnaire. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 48(11), 1238-1242. Sage CA: Los Angeles, CA: SAGE Publications.

Helton, W. S., & Näswall, K. (2015). Short Stress State Questionnaire: Factor structure and state change assessment. *European Journal of Psychological Assessment*, 31(1), 20.

Jamieson, G. A., & Skraaning Jr, G. (2018). Levels of automation in Human Factors models for automation design: Why we might consider throwing the baby out with the bathwater. *Journal of Cognitive Engineering and Decision Making*, 12(1), 42-49.

Jipp, M., & Ackerman, P. L. (2016). The Impact of Higher Levels of Automation on Performance and Situation Awareness. Journal of Cognitive Engineering and Decision Making, 10(2), 138–166.

Jipp, M. (2016). Expertise Development With Different Types of Automation. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 58(1), 92–106.

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, 5(2), 113–153.

Kaber, D. B. (2018). Issues in Human–Automation Interaction Modeling: Presumptive Aspects of Frameworks of Types and Levels of Automation. *Journal of Cognitive Engineering and Decision Making*, 12(1), 7–24.

Karpinsky, N. D., Chancey, E. T., Palmer, D. B., & Yamani, Y. (2018). Automation trust and attention allocation in multitasking workspace. *Applied Ergonomics*, 70, 194–201.

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, 1-15.

Körber, M., Weißgerber, T., Kalb, L., Blaschke, C., & Farid, M. (2015). *Prediction of take-over time in highly automated driving by two psychometric tests*, *82*(193), 195–201.

Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs (?). *Behavior Research Methods*, 51(1), 40-60.

Marsh, S., & Dibben, M. R. (2003). The role of trust in information science and technology. *Annual Review of Information Science and Technology (ARIST),* 37, 465-98.

Matthews, G., Joyner, L., Gilliland, K., Campbell, S., Falconer, S., & Huggins, J. (1999). Validation of a comprehensive stress state questionnaire: Towards a state big three. *Personality psychology in Europe*, 7, 335-350.

Matthews, G., De Winter, J., & Hancock, P. A. (2020). What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. *Theoretical issues in ergonomics science,* 21(4), 369-396.

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87.

McGarry, K., Rovira, E., & Parasuraman, R. (2003). Effects of task duration and type of automation support on human performance and stress in a simulated battlefield engagement task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 47(3), 548–552.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, *55*(3), 520-534.

Mogford, R. H., Guttman, J. A., Morrow, S. L., & Kopardekar, P. (1995). *The Complexity Construct in Air Traffic Control: A Review and Synthesis of the Literature.*

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of experimental psychology: Applied*, 6(1), 44.

Niu, J., Geng, H., Zhang, Y., & Du, X. (2018). Relationship between automation trust and operator performance for the novice and expert in spacecraft rendezvous and docking (RVD). Applied Ergonomics, 71(August 2017), 1–8.

Ockerman, J. J., & Pritchett, A. R. (2002, September). Impact of contextual information on automation brittleness. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 46(3), 382-386. Sage CA: Los Angeles, CA: SAGE Publications.

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors*, 56, 476–488.

Onnasch, L., Ruff, S., & Manzey, D. (2014). Operators′ adaptation to imperfect automation – Impact of miss-prone alarm systems on attention allocation and performance. *International Journal of Human-Computer Studies,* 72(10–11), 772–782.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, *30*(3), 286-297.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of cognitive engineering and decision making*, 2(2), 140-160.

R Development Core Team (2015) R: *A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink?. *Journal of experimental psychology: Human perception and performance*, 18(3), 849.

Redick, B. T. S., Shipstead, Z., Meier, M. E., Montroy, J. J., Hicks, K. L., Unsworth, N., … Hambrick, D. Z. (2016). Cognitive predictors of a common multi-tasking ability: Contributions from working memory, attention control, and fluid intelligence. *Journal of experimental psychology: General*, 145(11), 1473.

Roth, E. M., Sushereba, C., Militello, L. G., Diiulio, J., & Ernst, K. (2019). Function Allocation Considerations in the Era of Human Autonomy Teaming. *Journal of Cognitive Engineering and Decision Making*, 13(4), 199–220.

Rovira, E., Zinni, M., & Parasuraman, R. (2002a). Effects of Information and Decision Automation on Multi-Task Performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 46(3), 327–331.

Rovira, E., McGany, K., & Parasuraman, R. (2002b). Effects of unreliable automation on decision making in command and control. *In Proceedings of the Annual Meeting of the Human Factors and Ergonomics Society*, 428-432, Santa Monica, CA: Human Factors Society

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(1), 76–87.

Rovira, E., Pak, R., & McLaughlin, A. (2017). Effects of individual differences in working memory on performance and trust with various degrees of automation. *Theoretical Issues in Ergonomics Science*, 18(6), 573–591.

Sauer, J., Kao, C.-S., & Wastell, D. (2012). A comparison of adaptive and adaptable automation under different levels of environmental stress. *Ergonomics*, 55(8), 840–853.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., ... & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. Methods in Ecology and Evolution, 11(9), 1141-1152.Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.

Schumacher, E. H., Seymour, T. L., Glass, J. M., Fencsik, D. E., Lauber, E. J., Kieras, D. E., & Meyer, D. E. (2001). Virtually Perfect Time Sharing in Dual-Task Performance: Uncorking the Central Cognitive Bottleneck. *Psychological Science*, 12(2), 101–108.

Strand, N., Nilsson, J., Karlsson, I. C. M., & Nilsson, L. (2014). Semi-automated versus highly automated driving in critical situations caused by automation failures. *Transportation research part F: traffic psychology and behaviour*, 27, 218-228

Strybel, T. Z., Vu, K. P. L., Chiappe, D. L., Morgan, C. A., Morales, G., & Battiste, V. (2016). Effects of NextGen Concepts of Operation for Separation Assurance and Interval Management on Air Traffic Controller Situation Awareness, Workload, and Performance. *International Journal of Aviation Psychology*, 26(1–2), 1–14.

Trapsilawati, F., Wickens, C., Chen, C. H., & Qu, X. (2017). Transparency and Conflict Resolution Automation Reliability in Air Traffic Control. In *19th International Symposium on Aviation Psychology*, 419.

Ulrich, R., & Miller, J. (2008). Response grouping in the psychological refractory period (PRP) paradigm: Models and contamination effects. *Cognitive Psychology*, 57(2), 75-121.

Van Acker, B. B., Parmentier, D. D., Vlerick, P., & Saldien, J. (2018). Understanding mental workload: from a clarifying concept analysis toward an implementable framework. *Cognition, Technology and Work*, *20*(3), 351–365.

Visser, T. A. W., Ohan, J. L., & Enns, J. T. (2015). Temporal cues derived from statistical patterns can overcome resource limitations in the attentional blink. *Attention, Perception, and Psychophysics*, *77*(5), 1585–1595.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. Human Factors, 50(3), 433-441.

Wickens, C. D., Mccarley, J. S., Alexander, A. L., Thomas, L. C., Ambinder, M., Zheng, S., & Field, M. (2005). Model of Pilot Error. *Contract*, January, 213.

Wickens, C. D., Clegg, B. A., Vieane, A. Z., & Sebok, A. L. (2015). Complacency and Automation Bias in the Use of Imperfect Automation. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 57(5), 728–739.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: a synthesis of the literature. *Theoretical Issues in Ergonomics Science*, 8(3), 201–212.

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. B. (2010). Stages and Levels of Automation: An Integrated Meta-analysis. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(4), 389–393.

Wright, J. L., Chen, J. Y. C., Barnes, (2018). Human – automation interaction for multiple robot control: the effect of varying automation assistance and individual differences on operator performance. *Ergonomics*, 0139, 1–13.

CHAPTER 5

General Discussion

Central aims and predictions of the thesis

The broad question addressed in this thesis is whether individual differences in cognitive abilities modulate the impact of automation on task performance, situation awareness and operator workload (collectively referred to hereafter as 'task outcomes'). The nature of automation (i.e., *what* tasks are automated and *how much* the operator does) has been well established in the literature as a modulating influence on task outcomes. Well-designed automation reduces cognitive demands (Saqer & Parasuraman, 2014) thus freeing-up cognitive resources so that operators do not become overwhelmed and avoid situations where they do not have the cognitive capacity to handle task requirements (e.g., Kaber & Endsley, 2004; Saqer & Parasuraman, 2014). However, what was not previously investigated is whether individual variation in a task-relevant cognitive ability, such as multi-tasking, could interact with automation to modulate task outcomes. Such findings would support previous assertions that cognitive abilities should be a key consideration for the implementation and future development of automation in the workplace (Kaber, 2018; Saqer & Parasuraman, 2014; Sethumadhavan, 2009).

In examining the interaction between multi-tasking ability and automation, the following predictions were made. First, it was expected that multi-tasking would influence task outcomes. This prediction was based on previous studies showing multi-tasking ability assists in rapidly switching between multiple tasks (Chérif et al., 2018; Rubinstein et al., 2001; Schumacher et al., 2001) and is related to performance in many work environments such as ATC (Sethumadhavan, 2009), driving (Körber et al., 2015), aviation (Strybel et al., 2016) and military Command and Control (J. Chen & Terrence, 2009). Multi-tasking is also a relevant ability to examine as it differs between individuals (Barron & Rose, 2017), and comprises decision making, working memory, and attentional control among other related executive processes (Hambrick et al., 2010; Redick et al., 2016). Second, it was expected that automation would benefit task outcomes, since automation which works as intended may reduce task load thereby lightening demands on cognitive processes, including multi-tasking, in comparison to manual task performance conditions (Kaber & Endsley, 2004; Onnasch et al., 2014). Lastly, it was expected that multi-tasking ability and automation would interact such that an operator with lesser multi-tasking ability should benefit proportionally more from automation than an operator with greater multi-tasking ability, as differences in automation outcomes as a function of cognitive ability have been found in other domains (Jipp & Ackerman, 2016; Wright et al., 2018).

These predictions were initially examined in Chapter 2, which tested whether multi-tasking modulated the benefit of low-degree automation (DOA; Wickens et al., 2010) applied to acceptance and hand-off tasks in the simulated ATC environment. Chapter 3 aimed to replicate and extend the findings of Chapter 2 by testing for the benefit of low-DOA on conflict detection – a complex task more representative of the type of tasks automation is applied to in real-world

situations. The methodology established in Chapter 2 was again used which provided an opportunity to re-examine the basic effects in the acceptance and hand-off tasks with an improved multi-tasking index. Additionally, Chapter 3 re-evaluated SA, with more sensitive queries, and examined workload outcomes with the addition of the NASA-TLX subjective workload measure.

Lastly, Chapter 4 applied the established methodology to examine if multi-tasking ability modulated the impact of imperfect automation on complex task performance. While perfectly reliable automation may benefit performance and reduce workload (Manzey et al., 2012), imperfect automation has the potential to impair performance, reduce SA, and increase workload when it fails (Körber et al., 2015; Strand et al., 2014). Chapter 4 therefore examined whether operators with poorer multi-tasking ability would experience commensurately greater costs when automation was imperfect than those with greater multi-tasking ability. It was reasoned that operators with poorer multi-tasking ability might take longer to notice automation failures and lack the skills to take back manual control compared to operators with greater multi-tasking ability.

Key findings

Multi-tasking ability can modulate the benefits of low-DOA for simple tasks

This thesis found considerable evidence that poorer multi-taskers performed worse in the acceptance and hand-off tasks than better multi-taskers (Chapters 2 and 3). Poorer multi-tasking ability also impaired conflict detection performance (Chapters 2 and 4). These findings relating to individual differences were largely as expected, given multi-tasking is required for the ATC task and the measure of individual differences in multi-tasking was consistent across chapters. Low-DOA also had a consistent benefit for acceptance and hand-off tasks in all chapters, and for conflict detection when automated in Chapters 3 and 4. These findings relating to automation were expected given limited previous literature that low-DOA has been shown to improve task performance in other domains above performance with no automation (Manzey et al., 2012; Wright et al., 2018).

Crucially, multi-tasking ability also interacted with low-DOA in Chapters 2 and 3, such that poorer multi-taskers performed proportionally better on acceptance and hand-off tasks than better multi-taskers when assisted by low-DOA. However, multi-tasking ability did not modulate the low-DOA benefit to conflict detection in Chapters 3 or 4 or interact with low-DOA. Compared to the finding for the simpler acceptance and hand-off tasks, the conflict detection finding indicates that task complexity can interact with automation and cognitive ability to determine performance and is therefore an important factor to consider in deploying automation.

The key finding that multi-tasking ability modulated the benefit of low-DOA supports previous findings that individual differences play an important role in determining the outcomes for operators using automated systems (Jipp & Ackerman, 2016; Wright et al., 2018). Further, the

automation benefit found for both simple (acceptance and hand-offs) and complex (conflict detection) tasks, even with a relatively low-DOA, adds to the general body of literature which finds automated aids can have significant benefits over manual performance in simulated complex task environments (Chen et al., 2018; Kaber & Endsley, 2004; Manzey et al., 2012; Parasuraman et al., 2009), including in simulated ATC (Leiden et al., 2003; Trapsilawati et al., 2016).

Multi-tasking ability and low-DOA predicts situation awareness

Across all three experimental chapters, multi-tasking ability reliably predicted SA as expected. The size of the multi-tasking effect was similar (Chapter 2 $\beta = -0.28$, Chapter 3 $\beta = -0.25$, Chapter 4 $\beta = -0.16$), with poorer multi-taskers showing worse SA. However, low-DOA benefited SA only in Chapter 4, and only modestly ($\beta = -0.07$). Lastly, multi-tasking ability did not modulate the benefit of low-DOA for SA in any Chapter. The inconsistent benefits of low-DOA across chapters mirror the current literature which has shown a mixture of null results (Chen et al., 2017; Deng et al., 2019; Wright et al., 2018), automation use impairment to SA (Strybel et al., 2016), and automation use benefits to SA (e.g., Kaber & Endsley, 2004; Parasuraman et al., 2009).

The absence of reliable automation effects on SA here is puzzling. In this thesis, it is not the case that the SPAM measure was simply insensitive as demonstrated by its consistent relationship with variations in multi-tasking ability. One possibility is that SA benefits occur with some types of automation but not others. Studies which have found automation benefitted SA did so with adaptive automation which could be changed by the operator in a UAV task (Parasuraman et al., 2009) or with ten levels of automation that operators could choose between in an air traffic-type task (Kaber & Endsley, 2004). This more flexible automation may have encouraged participants to engage more with task-related events in order to know when is optimal to choose a different automation level (Kaber & Endsley, 2004) or engage the automation (Parasuraman et al., 2009) compared to participants here who used static low-DOA (i.e., present throughout the scenario at same DOA).

Another possibility is that SA simply may be a harder concept to examine in the context of automation in ATC than is often acknowledged in the literature. There are three potential reasons, which require further exploration. First, as described in Chapter 1, a more sensitive SA measure may be required to detect automation effects. It has been suggested that to find effects with objective freeze measures, 60 queries per condition are recommended (Endsley et al., 2000), which was not practical for the current studies with shorter scenarios which had 18 per scenario at roughly 90 second intervals. Second, as discussed in Chapter 3, the queries may have assessed task-relevant information that was nevertheless not related to what automation assisted with (i.e., highlighting when aircraft needed action). Third, as discussed in Chapters 3 and 4 it may be that

the task did not require complex mental models to be formed by participants as all display information was visible during queries, as evidenced by accuracy > 89% in all chapters.

Low-DOA can benefit subjective workload

As expected, low-DOA reliably reduced subjective workload compared to manually performed scenarios on both the single-item workload measure and the well-validated NASA-TLX (Hart et al., 2006) subjective workload measures, both when low-DOA was reliable and when it was imperfect. This finding sensibly reflects robust reductions in task load arising from low-DOA which usually leads to concomitant benefits to workload (Matthews et al., 2020). The workload benefits of partially reliable low-DOA obtained in Chapter 4 differ from an earlier ATC study which gave low-DOA assistance to expert and junior air traffic controllers for conflict detection and found junior controllers experienced higher workload under imperfect automation than under manual conditions (Di Nocera et al., 2006). However, there is a key difference, in that junior air traffic controllers are highly trained and likely experienced the manual condition as their normal baseline experience, where-as the present thesis had novice undergraduate students for whom the manual condition was quite difficult, evidenced by higher subjective workload in manual. Trained professionals therefore may have experienced imperfect automation as imposing an additional workload, while novices considered the aid of even imperfect automation to reduce workload compared to the manual condition. On the other hand, contrary to expectations, multi-tasking ability did not influence either measure of subjective workload in any chapter and did not interact with low-DOA. The lack of effect of multi-tasking ability on subjective workload was not in line with expectations, as both multi-tasking and workload are subject to shared attentional and executive processes (Redick et al., 2016; Vidulich & Tsang, 2012).

Of interest, the results from the objective workload measure (using latency to SA queries) also show considerable divergence from the subjective measures outcomes. Better multi-tasking was related to reductions in objective workload in Chapter 2, with no effect of low-DOA or interaction, directly mirroring the result found for subjective workload in that chapter. However, this differed from Chapter 3 where no multi-tasking or low-DOA effects were detected. Taken together, the workload findings showed diverging effects of multi-tasking and low-DOA on workload, as well as divergence between outcomes obtained using subjective and objective workload measures. The latter finding aligns with a review of workload methodologies by Matthews et al. (2020) as it supports their assertions that workload is not a unitary concept, hence the divergence between measures observed both in that review and the present work. Additionally, Matthews and Hancock suggest that different workload measures may index somewhat different constructs – in particular, subjective workload may instead measure self-regulation or metacognition. This suggestion could also account for the failure to find the predicted relationship between multi-tasking ability and workload across chapters in this thesis.

Multi-tasking ability may vary performance costs of low-DOA errors

In Chapter 4, multi-tasking ability was predicted to interact with the type of trial (manual, automation reliable, automation failure) such that poorer multi-taskers would perform poorer on hand-offs, acceptances, and conflict detection when automation made an error compared to when automation worked reliably. In line with previous chapters, the effect of partially reliable low-DOA on SA and workload was also examined, noting that a benefit on these measures might still be possible with automation reliability at 70%.

The findings of Chapter 4 provide initial evidence that multi-tasking ability modulates costs to performance observed for both acceptances and hand-offs when considering the effect of trial type. Thus, poorer multi-taskers had a greater difference between their manual, reliable automation and automation error trial performance compared to better multi-taskers which lends initial empirical support to suggestions that individual differences in cognition can impact automation failure outcomes. Follow-up analysis which compared performance in two conditions directly (manual versus reliable automation; manual versus automation error) broadly supports the assertion that multi-tasking ability varied the cost of automation error (in total four out of six of the outcomes measures had found this pattern). This aligns well with previous findings that people can overcome automation errors and perform well with imperfect automation, suggesting that at least some operators have sufficient abilities to compensate for automation failure (Onnasch et al., 2014; Wickens & Dixon, 2007). Additionally, the combined analysis revealed a smaller, but significant, benefit of low-DOA in Chapter 4 (acceptance $\beta = -0.12$, hand-off $\beta = -0.11$) compared to when low-DOA was perfectly reliable in the preceding Chapters (acceptance $\beta > -0.33$, hand-off $\beta > -0.21$). This supports the continued benefit of imperfect automation above a certain level of reliability (Wickens & Dixon, 2007).

It was also found that when automation makes errors, performance may drop below manual. This is in line with previous findings by Rovira et al. (2002) who simulated multi-tasking using an engine system monitoring task and found under a manual condition people were better at detecting malfunctions of the gauges than under the 50% reliable automation (either low or high automation) condition. A possible explanation for why performance on automation error trials dropped below manual performance is that imperfect automation may add an additional cognitive burden of monitoring the automation which is correct some of the time. Thus, more cognitive resources may be needed when dealing with imperfect automation than when performing a task manually. A second possibility is that the change in automation reliability within the low-DOA condition may have made it difficult to establish a baseline expectation about automation reliability and thus to deploy optimal cognitive resources in response to automation failure. The first five minutes of the present ATC task had no failure events, and after that time acceptance and hand-off failures occurred sporadically. Thus, participants could have taken an extended period after failures began to establish how frequently they were occurring.

Past studies have noted that when automation reliability changes unpredictably, people struggle to adapt their monitoring of the system compared to when reliability remains consistent or is known prior (Di Nocera et al., 2005).

Limitations

Interpretation of the experiments in this thesis are subject to several limitations. First, the operationalisation of multi-tasking was based on latent factors methodology used to estimate the shared variance of cognitive tasks hypothesised to tap into multi-tasking ability. The cognitive tasks (PRP, dual task and attentional blink) were chosen based on prior research reporting that these tasks loaded together on a factor representing response selection (Bender et al., 2018), and their theoretical associations with processes such as task switching (Visser et al., 1999), attentional control (Kawahara et al., 2005), and working memory (Chun & Potter, 1995; Jolicœur & Dell'Acqua, 1999), all of which are central components of multi-tasking (Redick et al., 2016). However, the loading of these tasks onto a single latent factor was inconsistent. In Chapter 2 the AB did not load adequately with the PRP and Dual task conditions. This resulted in a latent factor based on two tasks, when three or more is advisable for a robust structure (Draheim et al., 2019).

That said, in Chapters 3 and 4, the AB did load with the PRP and dual task conditions (although loadings sizes were variable across chapters). The somewhat inconsistent loading of the AB task with the dual and PRP tasks may stem from some differences in the nature of the tasks. While the AB shares encoding and retrieval (i.e., working memory) processes with the dual and PRP tasks (Chun & Potter, 1995), performance may have relied less on task switching than these tasks because both targets were presented in the same spatial location, drawn from the same stimulus class, and required the same type of response (Visser et al., 1999). Additionally, the AB task does not require speeded responses, thus reducing demands on response selection. That said, the consistency of results across all three chapters suggests that both the two- and three-task latent factors captured similar variance, which are likely to reflect multi-tasking ability.

A second limitation concerns the objective workload measure. The present thesis used a modified SPAM procedure in which the task display was paused (as is common with SAGAT) to minimize the potential for SPAM prompts to interfere with task performance and SA. However, this may have had some unintended consequences on the use of SPAM ready prompt latency as a measure of objective workload. Previous studies using SPAM latency (Strybel et al., 2016) did not pause the task while SPAM queries were answered, thus creating a need to balance the workload of the task against the workload of responding to the SPAM query. In contrast, pausing the task may have incentivised heterogeneity in participants' approach to choosing when to respond to the ready prompt. Some participants may have wished to finish any acceptance, hand-off, or conflict detection task they were presently engaged in when a 'ready for a query' prompt appeared. Should this have been the case, then the use of SPAM latency as an objective workload

measure would be valid here. Essentially, this explanation posits that it is likely that poorer multi-taskers who experienced greater workload would prefer to finish their present task before doing an SA query because returning to an interrupted task would impose an additional cognitive load which such less able multi-taskers could ill afford. However, other participants may have used the SPAM pause time strategically to give themselves a break, temporarily reduce their workload and reorient themselves while answering the SA query. These participants might thus be more likely to try to respond to the 'ready prompt' as soon as it appeared, particularly when workload was greater. For such participants faster responds could thus indicate higher workload, rather than slower responses, as is generally assumed to be the case in previous studies.

While this second approach to responding may be possible, however, evidence in favour of this option appears scant. Results from all chapters indicated that poorer multi-tasking ability was associated with longer RTs to 'ready' prompts, although this only reached statistical significance in Chapter 2. This would be expected if participants with poorer multi-tasking ability were experiencing greater workload and thus responding more slowly to 'ready' prompts. Further, examination of the distribution of 'ready' prompt RTs across participants suggests they were unimodally distributed. This is inconsistent with the idea two subgroups of participants developed different strategies for responding to the 'ready' prompts when experiencing higher workload, as this would be expected to generate a bimodal distribution with one peak at shorter RTs and another at longer RTs, reflecting the two different approaches.

A third potential limitation is the generalisability of the present findings to human-automation teaming in the workplace. The present sample of university-aged students were novices to the ATC task. Whereas novices are more likely to make mistakes and be involved in accidents, experts using automation in the workplace typically have significant experience and training, which benefits performance (Balfe et al., 2014; Jamieson & Skraaning, 2020). Expertise also can mitigate problems when automation fails, as experts can recover performance more efficiently (Roth et al., 2019). In addition, experts may have different perceptions of workload (Matthews et al., 2020), and their experience may mediate SA outcomes (Endsley, 2020; Cak et al., 2020). An illustrative example comes from Di Nocera et al. (2006) who compared expert and trainees air traffic controllers on conflict detection. They found the use of automation improved trainee performance compared to manual control, resulted in higher workload due to combined load of task and unfamiliar automation, and engendered greater trust in the system (Di Nocera et al., 2006). By comparison, experts' performance was unchanged from manual to automation, they had lower workload than trainees overall, and showed less trust in automation. Experts also recalled aircraft IDs better in a prospective memory task, suggesting they might also have had greater SA than novices. In sum, experts may have very different outcomes to novices, and thus the present findings with novices may not generalise to automation operators in the workplace (see Simons et al., 2017 for a current discussion of constraints on generality, also Arnett, 2008).

Another limitation of the experiments in this thesis is one inherent to many experiments conducted in university psychology departments; namely, the potentially limited generalisability of the findings to people outside the sample group. Since an early review in 1969 by Schultz, the issue of how representative psychology samples are of the general population has been investigated and debated. Current reviews have found up to 90% of participants in the highest impact developmental psychology journals come from a small selection of western countries (Neilsen et al., 2017; also see Pollet & Saxon, 2019). University samples in psychology often conform to a common set of ethnographic and sociological features described by the acronym WEIRD, being western, educated, industrialised, rich, and democratic. Relying on such samples is often considered problematic as a limited picture of humanity (see Apicella et al., 2020 for a current overview, also Henrich et al., 2010). This is further exacerbated when students differ from the general population, in factors such as ethnicity, work experience, education level, and socio-economic background (Sears, 1986). For the present results, this could mean the sample had higher multi-tasking skills due to the cultural focus put on multi-tasking, particularly in students who study across several units, and balance work, study, and social life. It is a cultural norm in western education institutions to place high demands on students, for instance delivering content across modalities (i.e., in-person, online) and assessments across communication styles (i.e., verbal presentations, written), which results in fast-paced and variable study conditions requiring considerable multi-tasking to navigate. Thus, in a non-student or non-WEIRD sample, the distribution of multi-tasking ability may be wider. This may change the strength of the multi-tasking effects; potentially reducing the strength of the effect because the effect may be small (as indicated in this thesis) and distributed over a greater range.

Particularly relevant to the present research, studies have shown cross-cultural differences in cognitive processes that are relevant to multi-tasking and the use of automation. For example, attention to visual information may vary across cultures, with a difference noted between central focus predominant in Western (typically individualist) cultures (Nisbett & Masuda, 2003) and a greater context-dependent (background) focus observed in Eastern (typically collectivist) cultures. This may be reflected in evidence that individuals from collectivist cultures tend to make more saccadic eye movements during visual search (Chua et al., 2005). Thus, to the extent that attentional and perceptual abilities that would influence multi-tasking may differ between cultures, there is some chance that the present findings may not be representative of automation use or interactions with cognition in non-WEIRD countries.

The sample was also novices. As mentioned in discussions of expertise, this may have resulted in a wider distribution of multi-tasking ability than could be seen with an expert sample. Expertise can be created by selecting people with superior cognitive abilities for a job, or with a natural predisposition to a job given personal strengths. Alternatively, expertise can be moulded

through experience, training, and interaction with other experts on the job. Thus, an expert sample could have a narrower distribution of high-level multi-tasking ability. This may, in turn, moderate the strength of the multi-tasking effect on automation.

Future directions

In recent times, many areas of research, both within and outside psychology, have turned to personalisation as a solution to problems. Often, automation is central to these personalised solutions. For instance, the use of automation such as smart robots to improve quality of elderly care (Breman, 2021), monitoring of multiple patients in hospital setpreferences, head-worn displays (Klueber et al., 2019), use of technology to personalise learning for engineering students (Svetsky et al., 2010) and in transportation particularly, where there has been a large growth in automation, research considering 'driver state, personal preferences and predictions of conditions' (Diederichs et al., 2020; 2022; Pangiltlpoulos & Dimitrakopoulous, 2022).

From a psychology perspective, it is important to consider how individual differences in cognition may play a role in developing such automated systems in the future. With this broad aim in mind, the following future directions aim to expand on the current research which focused on the interaction between individual differences and low-DOA automation in a predictable and relatively slow-paced task environment. This work will be critical to further our understanding of the boundary conditions of the current findings (see Sanderson & Burns, 2017 for discussion of such limits in empirical research) and thus how they can inform development of personalised automation solutions.

Investigating a different cognitive ability

The current thesis focused on multi-tasking because it is relevant to a broad set of domains, including transport, manufacture, military, and system monitoring in engineering. Multi-tasking is associated with individual differences across the population (Hambrick et al., 2009; Redick et al., 2016) and can be assessed readily using simulated complex task domains (e.g., ATC) or measuring the underlying faculties using cognitive tasks (e.g., working memory; Redick et al., 2016). Additionally, multi-tasking is directly aided by technology and automation in many workplace environments (Cullen et al., 2014). Another particularly relevant facet of multi-tasking is that is often volitional (i.e., comprises processes under the control of the operator) and a high-level cognitive skill based on several more elementary skills (e.g., direct, switch, or focus attention).

This last point begs the question of whether individual differences in less volitional processes would also modulate automation. For an alternative candidate process to be examined in the methodological framework set out in this thesis such a candidate process would need to be: 1) associated with individual differences, 2) capable of being indexed with brief, reliable and

valid cognitive tasks, 3) broadly relevant to workplace environments which involve automation, and 4) potentially aided by automation. With these conditions in mind, one potentially fruitful area for future experiments could be an examination of lower-level perceptual processes. Such processes would be clearly differentiated from the higher-order multi-tasking ability explored in this thesis, but still relevant to task performance, SA, and workload. One candidate process that meets this description is object tracking.

Over thirty years ago, Pylyshyn and Storm (1988) demonstrated that the human visual system can simultaneously track multiple targets across time and space. Object tracking involves preconceptual (i.e., early-stage) perceptual processes in early vision that connects objects and their visual representation in the mind (see Meyerhoff & Papenmeier, 2020 for review). Object tracking may be performed best using peripheral vision (Fehd & Seiffert, 2008, 2010). This phenomenon has also been studied by measuring brain region activation (McMains & Somers, 2004) with the posterior parietal cortex being posited as a potential candidate (Culham et al., 1998). Thus, object tracking relies on different processes and earlier stages of perception then multi-tasking.

Object tracking ability has also been demonstrated to vary between individuals. Pylyshyn suggested that people could track five objects over the short time of the task, while follow-up experiments have suggested greater variability with individuals being able to track as few as two, and as many as six objects (Oskama & Hyönä, 2004). Further, individual difference factors have been found to impact capacity, such as expertise (Allen et al., 2004) and video game experience (Green & Bavelier, 2007). Higher fluid intelligence has also been associated with higher capacity compared to verbal intelligence (Tullo et al., 2018). Thus, the capacity to track objects may be flexible rather than fixed and capable of expanding to meet requirements if assisted (Cavanagh & Alvarez, 2005).

Pylyshyn's multiple object tracking (MOT) paradigm is relatively brief, consisting of seven to 15 second trials where several items are initially presented on the screen, with a subset subsequently highlighted as targets. The target highlighting is then removed, and all items move randomly for a period of time and then stop. One of the items is highlighted again and observers must indicate whether it was one of the original targets or not. The MOT paradigm is a standardised task to index object tracking ability, well validated in a series of studies by Howe et al., (2010), shows high test-retest reliability (Meyerhoff & Papenmeier, 2020) and thus could be used in a future study in the same way the PRP and Dual tasks were presently used to index multi-tasking ability.

The capacity to track multiple objects is used in many aspects of everyday life; from driving, to watching team sports, and keeping an eye on your children in the playground (Alvarez & Franconeri, 2007). Object tracking is also relevant in many work domains which use automation, including the military, ATC, and system monitoring in engineering and transport.

Equally, object tracking capacity appears to have several properties that make it amenable to automated assistance. For example, the number of objects an individual is capable of tracking can be increased by reducing the speed or duration of tracking (Oskama & Hyönä, 2004), or reducing the number of distractor objects or the proximity of distractors (Cavanagh & Alvarez, 2005). Automation, such as low-DOA, could accomplish this by first identifying relevant targets based on classification criteria and highlighting them, thus reducing visual tracking requirements. This could be supplemented by higher DOA that makes decisions based on target identification. A real-world example of this may be in air traffic control, where higher DOA tracks aircraft and classifies them in terms of their likelihood of collision with other aircraft based on their speed, altitude, and flight path. If an aircraft deviates from its altitude or flight path, which could potentially lead to a conflict with another aircraft, high-DOA could flag the aircraft for the air traffic controller and send alerts to the pilots.

Investigating different variations and gradations of automation

Thirdly, and most critical to the aim of personalised automation, future experiments should examine other types of automation using the present methodological framework to create a more fine-grained picture of how individual differences interact with specific automation functions to predict task performance. The current experiments have extensively examined low-DOA in the ATC environment. Low-DOA combines an early stage (i.e., information processing/ acquisition) and low-level (i.e., human performs all required manual actions and decision making) automation. While this has provided preliminary evidence that individual differences in cognitive ability interact with the presence of low-DOA to predict task performance, there is a large scope to expand these inquires to understand better what sort of automation is best for a given individual. High-DOA has frequently been examined in the literature when comparing to low-DOA (see reviews by Manzey et al., 2012; Onnasch et al., 2014; Wright et al., 2018) and often improves performance, but involves significant complacency and risks if automation fails. For that reason, high-DOA is often avoided in safety critical environments. Therefore, intermediate DOAs are valuable to examine to fill in the gaps in the continuum of automation. Some studies vary the DOA, including medium-DOA (particularly when examining SA; Kaber & Endsley, 2004), but none examine these with regard to individual differences in cognitive ability to determine which operators may function best with such intermediate automation for a given task environment.

Medium-DOA has been conceptualised by Wickens et al., (2010) as stage 2 (information analysis) or stage 3 (action selection) at an intermediate level (allocating how much of a task is performed by the automation, and by the human). This middle part of the DOA continuum contains many gradations of stages and level combinations which could be explored to better understand how their benefits may vary with operator cognitive capabilities. For instance, a

medium-DOA narrows down options for the operator who is still required to make a final selection and execute the action. In ATC this could be implemented by the system highlighting only pairs of aircraft that will conflict in the near future and asking the operator whether the system should intervene by contacting the aircraft to change altitude. Future experiments could take a fine-grained approach by varying either the level or stage of automation in a task while keeping the other fixed. For example, a fixed low-level automation in which the human operator always performs the manual action, could be combined with early stage (i.e., information processing) automation that could highlight all information (e.g., incoming aircraft), medium stage automation (i.e., information acquisition or decision selection) that could only highlight critical information, or late-stage automation that presents possible actions to an operator for them to choose between. Such an approach, in combination with an examination of individual differences in perceptual and cognitive abilities, support finding the optimal amount or type of automation for any given individual.

Investigating a different task environment

Related to the question of generalisability comes a second question, 'do the patterns of effects found here occur in task domains other than ATC?' It is necessary to demonstrate a comparable pattern of findings using a different task to show broader generalisability and applicability, which may drive further investigation of the potential for personalised automation. The simulated environment used in this thesis was the ATC. This was chosen for several reasons: it required multi-tasking, it was flexible, as there were many ways of customising automation to assist multi-tasking separately for each aspect of the task, and it simulated a real-world environment while also broadly representing the types of jobs to which automation is applied. However, while this may be the case, ATC is not an environment that many people have experience with, and the slow-speed and predictability (e.g., fixed flight levels) of the ATC task used here does not reflect more dynamic settings, such as piloting an aircraft or driving a motor vehicle.

Examination of the existing literature suggests many candidate alternatives such Command and Control tasks (Rovira et al., 2007), powerplant and space station control room simulators (Burns et al., 2008; Niu et al., 2018), and UAV (Calhoun et al., 2009) simulators. Driving also represents a particularly relevant task environment to investigate further for several reasons. From a technological point of view, advances in driver safety and autonomous driving technology are increasingly relevant (see Dikmen & Burns, 2016 for a study of Tesla driver experience; also see Endsley, 2017 for a naturalistic study of situation awareness in Tesla drivers). From a cognitive perspective, driving requires performing multiple tasks (checking mirrors, monitoring road conditions, navigating), each of which could have different automation applied to assist. From an ecological validity perspective, driving is an everyday activity which most people

have regular experience with and is central to many industries and occupations both directly (e.g., transport and defence) and indirectly (e.g., retail and manufacture). Finally, driving is fundamentally different to the ATC in terms of pace (i.e., ATC is slow paced, whereas driving can be fast paced with dynamic conditions), interruptions (i.e., driving can have many interruptions of different modalities, whereas ATC had few), and interactions with other humans (i.e., ATC had none, driving can involve interactions with other road users, or passengers). Therefore, generalising the current findings to another task environment would increase the applicability and robustness of the current framework for investigating if operator cognitive abilities can vary the benefits and costs of automation.

Expertise and performance

As noted above, experts have different performance, SA, and workload outcomes to novices when using automation in complex task environments (Cak et al., 2020; Jamieson & Skraaning, 2020; Matthews et al., 2020; Roth et al., 2019). Complex workplaces such as ATC engage experts who gain experience over many years and receive extensive training. Therefore, it is a valuable question to investigate whether experts have different requirements from personalised automation than novices. To this end, future studies could test actual air traffic controllers using the current methodology to examine if differences in multi-tasking ability within an expert group still modulates the effect of low-DOA. Such studies may assist the development of automation tailored to experts. In terms of the impact on the current findings, greater complexity may decrease the differences between good and poorer multi-taskers (reduce the range of performance) and increase the benefits of automation for all participants commensurate with the literature on high-DOA in moderately complex tasks. But such effects would always be contingent on the sample – experts or novices. If novices were given very high-fidelity simulator this could lead to cognitive and perceptual overload which may be compensated for by employing more simplified mental models. Under these conditions, automation may not improve performance as much as was seen in the present studies, as novice participants would have a low-level of performance due to the extent of increased task complexity.

While recruiting experts may increase the validity of these findings to the people most likely to benefit from them in the workplace, likewise more realism in the task environment may increase the generalisability and validity of these findings. The air traffic control task that was used is considered medium fidelity meaning it incorporates important ingredients from actual air traffic control but is missing others. For instance, a higher fidelity ATC might include multiple communication channels to simulate communication between controllers, as well as controllers and pilots. This would increase visual information and introduce auditory information, which could create competing demands across information channels and interference across modalities.

Other information typically needed by controllers such as weather and windspeed could also be provided, creating additional controller tasks, and competing visual information, thus increasing multi-tasking requirements further. Therefore, a higher fidelity or more dynamic ATC task would require more multi-tasking, and potentially could make the effect of multi-tasking and its interaction with automation stronger. Additionally, a high-fidelity simulation would allow examination of different cognitive abilities – for instance prospective memory or working memory, as more information is presented to participants providing new opportunities for examining these domains. Thus, multi-tasking may be more important in more realistic simulators as more tasks would be performed across modalities and goals, and thus the impact of individual differences in multi-tasking ability and their interaction with automation may be stronger in more realistic tasks.

Contribution and concluding thoughts

The contribution of this thesis lies in the scope of outcomes examined and the depth to which the questions were explored. The methodologies employed in this thesis to explore the questions raised are complex and well designed. By way of overview: each study investigates six dependent variables (performance × 3, SA, workload × 2) with two independent variables (multi-tasking ability, automation), accounting for the potential confounding/ contributing effects of three covariates (IQ, task order, and practice) using linear mixed modelling to account for within and between subject's effects simultaneously using both random slopes and intercepts in the models. Such models allowed this thesis to draw conclusions about the effects of automation at the sample level between individuals (i.e., across a continuum of cognitive ability) and the relative effect for a given individual in terms of their performance under manual and automated conditions.

The statistical findings are consistent between chapters and have high reliability. The results showed identical patterns of significance and very similar effects across acceptance and hand-off tasks within each chapter *and* across comparable chapters, presenting two layers of replicability. Additionally, measures were iterated and improved across the chapters, including validating the SA queries and adding a subjective workload measure to allow converging evidence which further supported initial conclusions. The findings for SA and objective workload outcomes were also very similar across chapters, and subjective workload findings were comparable between measures within each chapter, further demonstrating the robustness and reliability of these findings.

The key independent variable in this thesis (multi-tasking) was measured using variance shared from multiple cognitive tasks concatenated using latent factors methodology. This approach to operationalising cognitive ability is rarely used in applied psychology investigations (Bender et al., 2018; Redick et al., 2016). However, the present findings showed that multi-

tasking operationalised in this way could reliably predict complex task performance across a range of task demands, situation awareness and workload. Thus, a latent-factors approach may be useful in future applied studies as a new standard for measuring cognitive abilities rather than traditional single-task measures or lengthy task batteries – both of which have come under recent criticism (Barron & Rose, 2017).

The approach to automation in this thesis developed a novel framework not previously considered; namely, examining differing effects of automation *within* a single DOA. Previous literature only examined automation's effects compared across degrees (e.g., from low to high). Some experiments have compared manual to automated performance (Galster et al., 2002; Kaber & Endsley, 2004; Onnasch et al., 2014), but none investigated the possibility that an individual's particular cognitive abilities may result in differing effects of automation in one person compared to others. This thesis also defined automation benefit relative a no-automation baseline, consistent with calls for the need for clear comparisons to find valid main effects when using automation (as noted in; Onnasch et al., 2014; also see Kaber & Endsley, 2004). Thus, the new framework created here may contribute to new lines of inquiry in human-automation teaming research.

In conclusion, the present findings support the need to consider individual differences in operator capabilities in the future design of workplace systems. The multi-tasking ability of operators of automation is important to know as underlying ability determines not only complex task performance, but workload and SA which are interrelated with the safe operation of automation. Better multi-tasking ability has been demonstrated to result in better performance with reliable and imperfect automation, increased SA, and lower workload in an ATC task than poorer multi-tasking ability. However, this thesis suggests those with poorer muti-tasking ability can receive proportionally greater benefits from low-DOA compared to their performance without automation than those with higher multi-tasking ability. The proportionally greater benefits were only found for relatively simple tasks, but more complex tasks also benefited from low-DOA, although this effect was not varied by cognitive ability. These findings for low-DOA contribute to present understandings about the relative benefits of automation as these challenge traditional assumptions that higher DOAs are necessary to confer significant performance benefits. When accounting for operator's cognitive capabilities, this assumption may not be the case. Present findings relating to imperfect automation indicate there is a comparable cost to performance which is modulated by multi-tasking ability as was found for the benefits of automation. This strengthens the rationale for considering individual's cognitive abilities when determining what automation should be deployed in the workplace, as those who gained a benefit from reliable automation also suffered the greatest cost when automation made an error. Therefore, current design approaches that provide everyone with the same automated assistance should consider the potential benefits of a "many-sizes-fits-all" approach that provides automated assistance flexibly depending on capabilities and need.

References

Allen, R., McGeorge, P., Pearson, D., & Milne, A. B. (2004). Attention and expertise in multiple target tracking. *Applied Cognitive Psychology*, *18*(3), 337–347.

Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision*, *7*(13), 14.

Apicella, C., Norenzayan, A., & Henrich, J. (2020). Beyond WEIRD: A review of the last decade and a look ahead to the global laboratory of the future. *Evolution and Human Behavior, 41(5), 319-329.*

Arnett, J. (2008). The neglected 95%: Why American Psychology needs to become less American. *The American Psychologist*, 63, 602–614.

Balfe, N., Sharples, S., & Wilson, J. R. (2015). Impact of automation: Measurement of performance, workload, and behaviour in a complex control environment. *Applied Ergonomics*, *47*, 52–64.

Barron, L. G., & Rose, M. R. (2017). Multitasking as a Predictor of Pilot Performance: Validity Beyond Serial Single-Task Assessments. *Military Psychology*, *29*(4), 316–326.

Bender, A., Loft, S., Lipp, & Visser, T.A.W. (2018). Advancing our understanding of warfighter cognition: Development of a "cognitive profiling" tool to enhance situation awareness. *Defence Science and Technology (DST) Group Human Performance Research Network (HPRnet).*

Breman, P. (2021). How automation technology may contribute to quality of life in the elderly care: an integrated personalised care approach. *International Journal of Integrated Care, 21(S1).*

Burns, C. M., Skraaning Jr, G., Jamieson, G. A., Lau, N., Kwok, J., Welch, R., & Andresen, G. (2008). Evaluation of ecological interface design for nuclear process control: situation awareness effects. *Human Factors*, *50*(4), 663-679.

Cak, S., Say, B., & Misirlisoy, M. (2020). Effects of working memory, attention, and expertise on pilots' situation awareness. *Cognition, Technology and Work*, *22*(1), 85–94.

Calhoun, G., Draper, M., & Ruff, H. A. (2009). Effect of level of automation on unmanned aerial vehicle routing task. *Proceedings of the Human and Ergonomics Society Annual Meeting* 53(4), 197-201. Sage CA: Los Angeles, CA: Sage Publications.

Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences*, *9*(7), 349–354.

Chen, J. Y. C., & Terrence, P. I. (2009). Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment. *Ergonomics*, *52*(8), 907–920.

Chen, S. I., Visser, T. A. W., Huf, S., & Loft, S. (2017). Optimizing the balance between task automation and human manual control in simulated submarine track management. *Journal of Experimental Psychology: Applied*, *23*(3), 240–262.

Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, *19*(3), 259–282.

Chérif, L., Wood, V., Marois, A., Labonté, K., & Vachon, F. (2018). Multitasking in the military: Cognitive consequences and potential solutions. *Applied Cognitive Psychology*, *32*(4), 429–439.

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences, 102(35), 12629-12633.*

Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental psychology: Human perception and performance*, *21*(1), 109.

Culham, J. C., Brandt, S. A., Cavanagh, P., Kanwisher, N. G., Dale, A. M., & Tootell, R. B. (1998). Cortical fMRI activation produced by attentive tracking of moving targets. *Journal of Neurophysiology*, 80, 2657–70

Cullen, R. H., Dan, C. S., Rogers, W. A., & Fisk, A. D. (2014). The Effects of Experience and Strategy on Visual Attention Allocation in an Automated Multiple-Task Environment. *International Journal of Human-Computer Interaction*.

Deng, Y., Shirley, J., Zhang, W., Kim, N. Y., & Kaber, D. (2019). Influence of dynamic automation function allocations on operator situation awareness and workload in unmanned aerial vehicle control. *In International Conference on Applied Human Factors and Ergonomics*, 337-348. Springer, Cham.

Di Nocera, F.., Fabrizi, R., Terenzi, M., & Ferlazzo, F. (2006). Procedural errors in air traffic control: effects of traffic density, expertise, and automation. *Aviation, space, and environmental medicine*, 77(6), 639-643.

Di Nocera, F., Lorenz, B., & Parasuraman, R. (2005). Consequences of shifting from one level of automation to another: main effects and their stability. Human factors in design, safety, and management, 363-376.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation*, *14*(1), 20.

Diederichs, F., Knauss, A., Wilbrink, M., Lilis, Y., Chrysochoou, E., Anund, A., ... & Bischoff, S. (2020). Adaptive transitions for automation in cars, trucks, buses and motorcycles. *IET Intelligent Transport Systems, 14(8), 889-899.*

Diederichs, F., Wannemacher, C., Faller, F., Mikolajewski, M., Martin, M., Voit, M., ... & Piechnik, D. (2022). Artificial Intelligence for Adaptive, Responsive, and Level-Compliant Interaction in the Vehicle of the Future (KARLI). In *International Conference on Human-Computer Interaction, 164-171. Springer, Cham.*

Dikmen, M., & Burns, C. (2016). Autonomous Driving in the Real World: Experiences with Tesla Autopilot and Summon. *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '16),* Ann Arbor, MI, USA., May, 225–228.

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin*, *145*(5), 508.

Endsley, M. R., Sollenberger, R., & Stein, E. (2000). *Situation awareness: A comparison of measures. In Proceedings of the Human Performance, Situation Awareness and Automation: User-Centered Design for the New Millennium* (pp. 15–19). Savannah, GA: SA Technologies, Inc.

Endsley, M. R. (2017). Autonomous driving systems: A preliminary naturalistic study of the Tesla Model S. *Journal of Cognitive Engineering and Decision Making*, 11(3), 225-238.

Endsley, M. R. (2020). The divergence of objective and subjective situation awareness: A meta-analysis. Journal of cognitive engineering and decision making, 14(1), 34-53.

Fehd H. M. Seiffert A. E. (2008). Eye movements during multiple object tracking: Where do participants look? *Cognition*, 108, 201–209.

Fehd, H. M., & Seiffert, A. E. (2010). Looking at the center of the targets helps multiple object tracking. *Journal of vision*, *10(4), 19-19.*

Galster, S. M., Bolia, R. S., & Parasuraman, R. (2002, September). Effects of information automation and decision-aiding cueing on action implementation in a visual search task. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 46(3), 438-442. Sage CA: Los Angeles, CA: SAGE Publications.

Green, C. S., & Bavelier, D. (2007). Vision Action- Video-Game Experience Alters the Spatial Resolution of Vision. *Psychological Science*, *18*(1), 88–94.

Hambrick, D. Z., Oswald, F. L., Darowski, E. S., Rench, T. A., & Brou, R. (2010). Predictors of multitasking performance in a synthetic work paradigm. *Applied Cognitive Psychology*, *24*(8), 1149–1167.

Hart, S. G., & Wickens, C. D. (2010). Cognitive workload. *NASA human systems integration handbook*, 1-17.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and brain sciences*, *33*(2-3), 61-83.

Howe, P. D., Cohen, M. A., Pinto, Y., & Horowitz, T. S. (2010). Distinguishing between parallel and serial accounts of multiple object tracking. *Journal of Vision*, 10(8), 11-11.

Jamieson, G. A., & Skraaning, G. (2020). The absence of degree of automation trade-offs in complex work settings. *Human factors*, *62*(4), 516-529.

Jipp, M., & Ackerman, P. L. (2016). The Impact of Higher Levels of Automation on Performance and Situation Awareness. *Journal of Cognitive Engineering and Decision Making, 10*(2), 138–166.

Jolicœur, P., & Dell'Acqua, R. (1999). Attentional and structural constraints on visual encoding. *Psychological research*, 62(2), 154-164.

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*(2), 113–153.

Kawahara, J., Gabari, Y., & Enns, J. T. (2005). Testing the two-stage competition model of the attentional blink: Competition or a cost in distractor rejection?. *Journal of Vision*, 5(8), 112-112.

Klueber, S., Wolf, E., Grundgeiger, T., Brecknell, B., Mohamed, I., & Sanderson, P. (2019). Supporting multiple patient monitoring with head-worn displays and spearcons. *Applied ergonomics*, *78*, *86-96.*

Körber, M., Weißgerber, T., Kalb, L., Blaschke, C., & Farid, M. (2015). *Prediction of take-over time in highly automated driving by two psychometric tests. 82*(193), 195–201.

Leiden, K.J., Kopardekar, P., & Green, S., (2003). *Controller workload analysis methodology to predict Increases in airspace capacity.* Paper presented at the AIAA's 3rd Annual Aviation Technology, Integration, and Operations (ATIO) Tech, November 2003, Denver, Colorado (Reston, VA: AIAA).

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87.

Matthews, G., De Winter, J., & Hancock, P. A. (2020). What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. *Theoretical Issues in Ergonomics Science*, *21*(4), 369–396.

McMains, S. A., & Somers, D. C. (2004). Multiple spotlights of attentional selection in human visual cortex. *Neuron*, *42*(4), 677-686.

Meyerhoff, H. S., & Papenmeier, F. (2020). Individual differences in visual attention: A short, reliable, open-source, and multilingual test of multiple object tracking in PsychoPy. *Behavior Research Methods*, *52*(6), 2556-2566.

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, 162, 31-38.

Nisbett, R. E., & Masuda, T. (2003). Culture and point of view. *Proceedings of the National Academy of Sciences*, *100*(19), 11163-11170.

Niu, J., Geng, H., Zhang, Y., & Du, X. (2018). Relationship between automation trust and operator performance for the novice and expert in spacecraft rendezvous and docking (RVD). *Applied Ergonomics*, 71(August 2017), 1–8.

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human Performance Consequences of Stages and Levels of Automation. *Human Factors: The Journal of Human Factors and Ergonomics Society*, *56*(3), 476–488.

Oksama, L., & Hyönä, J. (2004). Is multiple object tracking carried out automatically by an early vision mechanism independent of higher-order cognition? An individual difference approach. *Visual Cognition*, 11, 631–671.

Orban, Panagiotopoulos, I., & Dimitrakopoulos, G. (2022). Cognitive selection of driving automation levels in highly automated vehicles leveraging on Bayesian networking principles. *IET Intelligent Transport Systems.*

Parasuraman, R., Cosenzo, K., & de Visser, E. (2009). Adaptive Automation for Human Supervision of Multiple Uninhabited Vehicles: Effects on Change Detection, Situation Awareness, and Mental Workload. *Military Psychology*, *21*(2), 270.

Pollet, T. V., & Saxton, T. K. (2019). How diverse are the samples used in the journals 'evolution & human behavior' and 'evolutionary psychology'?. *Evolutionary Psychological Science, 5(3), 357-368.*

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial vision*, *3*(3), 179-197.

Redick, B. T. S., Shipstead, Z., Meier, M. E., Montroy, J. J., Hicks, K. L., Unsworth, N., Kane, M. J., & Hambrick, D. Z. (2016). *Cognitive predictors of a common multitasking ability: Contributions from working memory, attention control, and fluid intelligence*.

Roth, E. M., Sushereba, C., Militello, L. G., Diiulio, J., & Ernst, K. (2019). Function Allocation Considerations in the Era of Human Autonomy Teaming. Journal of Cognitive Engineering and Decision Making, 13(4), 199–220.

Rovira, E., Zinni, M., & Parasuraman, R. (2002). *On multi-task performance*. *2000*, 327–331.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(1), 76–87.

Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive Control of Cognitive Processes in Task Switching. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(4), 763–797.

Sanderson, P., & Burns, C. (2017). Rasmussen and the boundaries of empirical evaluation. *Applied Ergonomics, 59, 649-656.*

Saqer, H., & Parasuraman, R. (2014). Individual performance markers and working memory predict supervisory control proficiency and effective use of adaptive automation. *International Journal of Human Factors and Ergonomics*, *3*(1), 15–31.

Schumacher, E. H., Seymour, T. L., Glass, J. M., Fencsik, D. E., Lauber, E. J., Kieras, D. E., & Meyer, D. E. (2001). Virtually Perfect Time Sharing in Dual-Task Performance: Uncorking the Central Cognitive Bottleneck. *Psychological Science*, *12*(2), 101–108.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of personality and social psychology*, *51*(3), 515.

Sethumadhavan, A. (2009). Effects of automation types on air traffic controller situation awareness and performance. *Proceedings of the human factors and ergonomics society 53rd annual meeting*, 2009(1), 1329–1333.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12(6), 1123-1128.*

Strand, N., Nilsson, J., Karlsson, I. C. M., & Nilsson, L. (2014). Semi-automated versus highly automated driving in critical situations caused by automation failures. *Transportation Research Part F: Traffic Psychology and Behaviour*, *27*, 218–228.

Strybel, T. Z., Vu, K. P. L., Chiappe, D. L., Morgan, C. A., Morales, G., & Battiste, V. (2016). Effects of NextGen Concepts of Operation for Separation Assurance and Interval Management on Air Traffic Controller Situation Awareness, Workload, and Performance. *International Journal of Aviation Psychology*, *26*(1–2), 1–14.

Svetsky, S., Moravcik, O., Sobrino, D. R., & Stefankova, J. (2010). The implementation of the personalised approach for technology enhanced learning. In *Proceedings of the World Congress on engineering and computer science (1).*

Trapsilawati, F., Chen, C. H., & Khoo, L. P. (2016). An investigation into conflict resolution and trajectory prediction AIDS for future air traffic control. *Advances in Transdisciplinary Engineering*, *4*, 503–512.

Tullo, D., Faubert, J., & Bertone, A. (2018). The characterization of attention resource capacity and its relationship with fluid reasoning intelligence: A multiple object tracking study. *Intelligence*, 69, 158-168.

Vidulich, M. A., & Tsang, P. S. (2012). Mental workload and situation awareness (chap. 8), and Salvendy, G. (Ed.). *Handbook of Human Factors and Ergonomics*.

Visser, T. A., Bischof, W. F., & Di Lollo, V. (1999). Attentional switching in spatial and nonspatial domains: Evidence from the attentional blink. *Psychological Bulletin*, 125(4), 458.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: a synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212.

Wickens, C. D., Mccarley, J. S., Alexander, A. L., Thomas, L. C., Ambinder, M., Zheng, S., & Field, M. (2005). Model of Pilot Error. *Contract*, January, 213.

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. B. (2010, September). Stages and levels of automation: An integrated meta-analysis. *In Proceedings of the human factors and ergonomics society annual meeting*, 54(4), 389-393. Sage CA: Los Angeles, CA: Sage Publications.

Wright, J. L., Chen, J. Y. C., & Barnes, M. J. (2018). Human-Automation Interaction for Multiple Robot Control: The Effect of Varying Automation Assistance and Individual Differences on Operator Performance. *Ergonomics*, 1–34.

Supplementary material

Chapter 2

On pages 64 - 65 are described the cognitive tasks. Here are the written instructions given to participants to read before completing each task.

Psychological Refractory Period Task:

Task instructions: The experiment is looking at how quickly you can respond to symbols and sounds.

Practice Phase

Symbol Practice: First you will be shown on the computer screen 4 symbols and the corresponding response keys. Please ensure to memorize the response mappings before pressing the space key to move on to the practice trials.

Sound Practice: After the symbol practice, you will be shown a screen with some plus signs and the corresponding response key under each plus sign.

Depending on the response mapping condition that you have been allocated to, the following screen may be shown:

+ + + +

A S D F

OR

+ + + +

H J K L

Each + sign represents a specific sound. In order to practice the response mappings, press the associated response keys on your keyboard and you will hear the sound relevant response key. Please ensure to practice the response mappings (by pressing on the specific keys) until you have memorized the correct mappings. Press the space bar only once you feel comfortable to start the actual practice phase.

Dual task Practice

Now things will get a little harder. You will be shown a symbol and a sound will be played either very closely after the symbol or with a slightly bigger delay between the symbol and the sound.

Try to respond to both as quickly and accurately as possible after seeing a symbol and hearing a sound**. It is important that you do not group your responses – i.e., make sure to respond to the first target first, followed by the second target.**

Testing Phase

This phase will be exactly the same as the dual task practice. You will be always shown a symbol and played a sound.

Try to respond to both as quickly and accurately as possible after seeing a symbol and hearing a sound**.**

Task Goal: Try your very best to answer as quickly and accurately to both, the symbol and sound.

Attentional Blink Task:

Task Instructions: For this task you will be shown a series of letters and numbers that are presented very rapidly in a serial manner. Your task is to identify two white target letters.

You will also see some numbers or abstract symbols in the presentation stream – these can be ignored.

At the end of each trial, you will be asked to report what the two white letters were. Here, it is important to report the target letters in the correct sequence – i.e., for Target 1 report the first letter that you saw in the stream and for Target 2, report the second letter in the stream.

When you have reported your response, you can press return to continue to the next trial.

Sometimes the letters will be quite easy to see, and other times more difficult. Please ensure to always report 2 target letters. If you are not sure, give it your best guess.

Task Goal: Try your very best to identify the two white letters that are presented in each stream.

Dual versus Single Response Selection Task:

Task instructions: The experiment is looking at how quickly you can respond to images and sounds.

Practice Phase

Shape Practice: First you will be shown (and asked to remember) 2 shapes and the corresponding response keys. Once you feel comfortable you will be given some practice.

Sound Practice: After practicing responding to the shapes, you will be shown (and asked to remember) 2 sounds and the corresponding response keys. Once you feel comfortable you will be given some practice.

Dual task Practice

Now things will get a little harder. On some trials, you will either be played a sound OR shown a shape. On other trials you will be played a sound AND shown a shape at the SAME time. Please respond as quickly and as accurately as you can.

If you are shown a shape and played a sound, respond to both as quickly and accurately as possible but please ensure to not press both response keys at the same time. Instead, press one of the response keys first, with the second response key quickly following the first.

Testing Phase

This phase will be exactly the same as the dual task practice, where you will either be played a sound OR shown a shape. On other trials you will see the shape AND hear a sound at the SAME time.

Task Goal: Try your very best to answer as quickly and accurately to either a single stimulus OR two stimuli together.

On page 65-66, I described the nature of the ATC task. In addition to the two conditions (manual and automated) there was also two experimental scenarios, each containing a roughly equivalent number of events. These were designated Scenario 1 and Scenario 3 (there was also a Scenario 2, which was used as the practice of the manual condition). The conditions and scenarios were counterbalanced, such that participants completed for instance manual (scenario 1) and low-DOA (scenario 3) or manual (scenario 3) and low-DOA (scenario 1). Thus, for each scenario there was a manual and an automated version. Below is a summary of when each acceptance, hand-off and conflict detection event occurred in each scenario, followed by the timings of those events.

Table 1: ATC event summary

Scenario 1

Aircraft ID	Acceptance Flashing Time Start	Max Acceptance Time	Handoff Flashing Time Start	Max Handoff	Max Conflict Time	Near miss Pair aircraft
QF73	0.13	0.33	6.9	7.10	3.24	
QF33	0	0	2.47	3.07		
AA97	0	0	1.2	1.40		
EK11	0	0	1.45	2.05		
NZ51	0	0	0.34	0.54		
VA23	0	0	3.28	3.48		
QF45	0	0	4.16	4.36		SQ48
SQ48	0	0	2.33	2.53		QF45
VA53	0	0.18	4.42	5.02	3.24	
EK18	0.38	0.58	6.03	6.23		
SQ57	2.12	2.32	6.48	7.08		
NZ29	2.33	2.53	7.43	8.03		
VA98	2.33	2.53	7.09	7.29		
QF94	3.14	3.34	8.21	8.41	7.08	
NZ31	4.02	4.22	8.59	9.19		
SQ77	4.21	4.41	9.03	9.23		
AA36	5.06	5.26	10.56	11.16	7.08	
EK30	5.18	5.38	10.31	10.51		NZ20
NZ20	6.37	6.57	12.02	12.22		EK30
QF59	6.41	7.01	12.37	12.57	10.11	
SQ27	7.05	7.25	11.56	12.16	10.11	
AA58	8.14	8.34	14.15	14.35	13.08	
NZ17	9.07	9.27	14.20	14.40	13.08	
QF89	9.43	10.03	15.54	16.14	14.43	
EK56	10.23	10.43	16.38	16.58	14.43	
QF52	10.42	11.02	15.57	16.17		AA35
AA34	11.11	11.31	17.10	17.30		
QF10	12.34	12.54	18.13	18.33		
NZ16	13.08	13.28	18.21	18.41		
AA35	13.34	13.54	19.28	19.48		QF41
VA12	13.24	13.44	18.20	18.40		QF52
QF41	14.26	14.46	19.52	20.12		VA12
SQ47	15.07	15.27	20.50	21.10	17.10	
VA37	15.04	15.24	20.00	20.20	17.10	
QF87	16.28	16.48	22.41	23.01	21.35	
EK81	17.80	18.00	23.05	23.25	21.35	
AA63	17.42	18.02	23.43	24.03		NZ19
QF71	18.13	18.33	24.11	24.31		
NZ19	19.15	19.35	24.19	24.39		AA63
AA96	20.05	20.25	26.07	26.27	23.46	

QF84	20.12	20.32	27.35	27.55		NZ40
VA14	21.02	21.22	25.49	26.09	23.46	
EK61	21.36	21.56	26.55	27.15		
QF76	22.24	22.44	28.41	29.01	26.08	
AA79	23.04	23.24	28.48	29.08	26.08	
QF55	23.52	24.12	30.01	30.21		
QF75	24.25	24.45	29.45	30.05	28.38	
NZ40	25.01	25.21	29.57	30.17		QF84
AA42	26.15	26.35		-	28.38	
QF32	26.15	26.35		-		
SQ70	27.36	27.56		-		
AA90	28.09	28.29		-		
VA91	29.18	29.38		-		

Scenario 3

Aircraft ID	Acceptance Flashing Time Start	Max Acceptance Time	Handoff Flashing Time Start	Max Handoff	Max Conflict Time	Near miss pair aircraft
QF97	0	0	4.45	5.05		SQ57
QF30	0	0	0.56	1.16		
NZ89	0	0	0.01	0.21		
Ek23	0	0	4.33	4.53		
SQ57	0	0	3.22	3.42		QF97
AA31	0	0	2.50	3.10		
AA64	0	0	2.19	2.39		
QF24	0	0	1.42	2.02		
AA16	0.29	0.49	6.16	6.36	4.06	
VA13	0.44	1.04	6.26	6.46	4.06	
EK26	1.33	1.53	6.53	7.13		
QF52	2.20	2.40	7.59	8.19		
AA37	2.36	2.56	8.36	8.56		NZ45
QF62	3.06	3.26	9.12	9.32	6.44	
AA41	3.22	3.42	9.47	10.07	6.44	
NZ45	3.02	3.22	9.40	10.00		AA37
EK53	5.09	5.29	10.10	10.30		
QF70	5.26	5.46	10.43	11.03		AA35
VA32	6.44	7.04	12.29	12.49	8.55	
SQ79	7.11	7.31	12.03	12.23	8.55	QF70
AA35	8.21	8.41	13.53	14.13		
Ek17	9.05	9.25	14.27	14.47	11.45	
SQ73	9.12	9.32	14.34	14.54		
VA29	10.03	10.23	14.56	15.16	11.45	
AA83	10.19	10.39	16.34	16.54	15.22	
NZ65	10.35	10.55	15.46	16.06	15.22	
NZ55	11.25	11.45	16.45	17.05		
QF40	0.32	0.52	18.31	18.51		NZ94

AA48	12.58	13.18	20.34	20.54		
VA71	0.37	0.57	18.25	18.45	17	
QF12	13.59	14.19	20.05	20.25	17	
QF42	15.09	15.29	21.42	22.02		
NZ94	15.55	16.15	21.33	21.53		QF40
QF86	16.16	16.36	21.19	21.39	20.12	
QF59	17.08	17.28	23.11	23.31	22.02	
EK85	18.03	18.23	23.40	24.00	22.02	
AA38	18.15	18.35	24.25	24.45	20.12	
QF20	18.58	19.18	25.20	25.40		
QF58	19.30	19.5	25.24	25.44		
VA28	20.34	20.54	25.13	25.33		SQ56
SQ56	20.55	21.15	26.03	26.23		VA28
EK60	21.26	21.46	26.58	27.18		NZ84
Ek88	22.24	22.44	28.40	29.00	26.08	
NZ84	22.53	23.13	28.14	28.34		EK60
E99	23.03	23.23	28.48	29.08	26.09	
QF91	24.33	24.53	29.55	30.15		
SQ51	25.37	25.57	0			
QF81	25.50	26.10	0			
NZ33	26.12	26.32	28.35	28.55		
QF66	27.16	27.36	0			
VA95	27.16	27.36	28.35	28.55		
VA74	28.16	28.36	0			

In chapter 2 I describe the linear mixed models used to test the hypotheses. In the first instance, model comparisons are presented here show the comparison between a null model containing only the intercept and random effects, and the interaction model for each dependent variable. This demonstrates that models with the interaction of condition and multi-tasking were the models with the greatest predictive power.

Null model formula: dependent variable ~ (1 | participant_number) + (0 + condition | participant_number)

*Interaction model formula: dependent variable ~ condition * MT_Factor + (1 | participant_number) + (0 + condition | participant_number)*

Table 2: Model comparisons between null model and full model for each dependent variable model using Chi-square test of significance.

Model	AIC	BIC	Log Likelihood	Deviance	Chisq	Df	Significance
Acceptances Null	44020	44048	-22006	44012			
Acceptances Model Full	43924	43974	-21955	43910	101.77	3	<.001
Hand-offs Null	45274	45302	-22633	45266			
Hand-offs Model Full	45218	45268	-22602	45204	61.65	3	<.001
Conflict Null	22426	22448	-11209	22418			
Conflict Model Full	22420	22459	-11203	22406	11.83	3	<.01
SA Null	23563	23587	-11777	23555			
SA Model Full	23537	23579	-11761	23523	31.84	3	<.001
Objective Workload Null	12326	12350	-6159.10	12318			
Objective workload Full	12324	12367	-6155.3	12310	7.64	3	0.05

Note: Log likelihood is a measure of fit describing how well a parameter explains the observed data. AIC is Akaike Information Criterion, an extimator of prediction error. Lower AIC reflects better model fit. BIC is Bayesian Information Criterion, a measure for model comparison or selection. Lower BIC reflect better model fit. Chisquare is a test of significant difference between models.

Additionally, models were conducted which included previous ATC experience, condition order, non-verbal intelligence and subjective rating of task importance were included as covariates.

*Formula: Dependent variable ~ condition * MT_Factor + IQ + Counterbalance + Experience + (1 | participant_number) + (0 + condition | participant_number)*

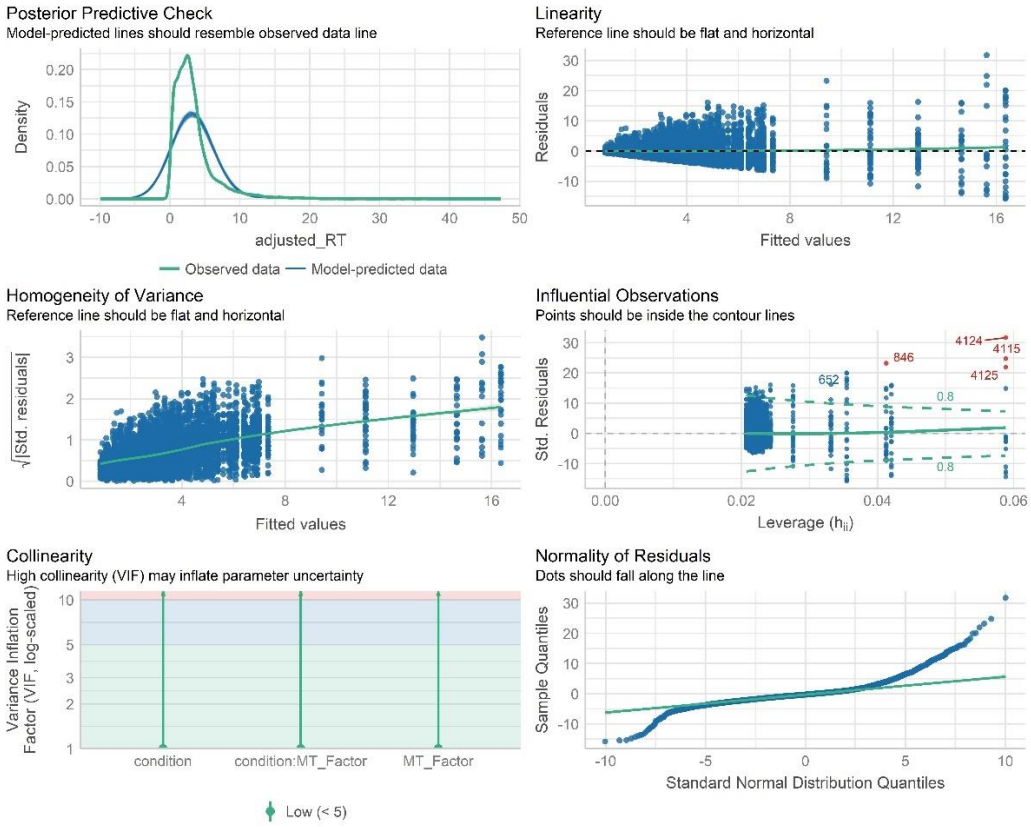
These were included to control for potential confounding effects. These covariate models are presented here. None of these covariates influenced the results of the main effects of interest.

Table 3: Covariate models unstandardised standardised regression effects. Standard deviation presented in parentheses.

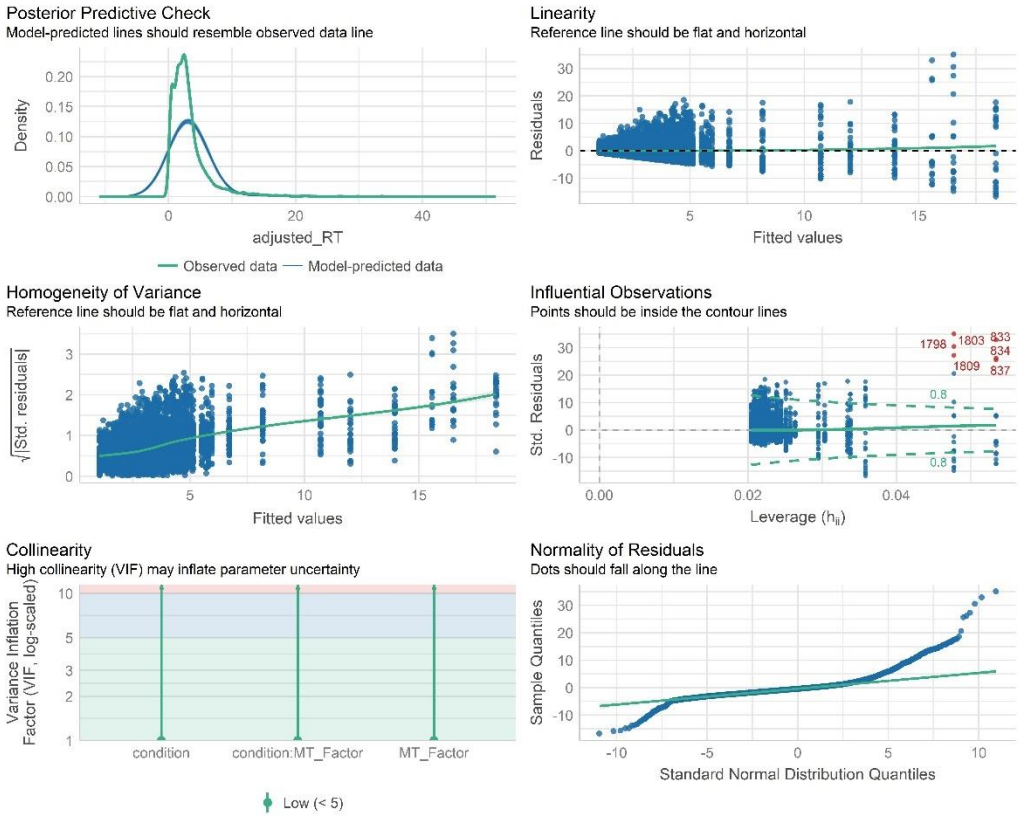
Parameter	Acceptances Adjusted RT	Hand-offs Adjusted RT	Conflict Adjusted RT	SA Question Adjusted RT	Objective Workload
Condition	-2.02 (0.19)***	-1.60 (0.22)***	-1.00 (3.35)	-0.04 (0.49)	-0.15 (0.08)
Multi-tasking	-0.85 (0.27)**	-0.94 (0.30)**	-8.57 (4.88) *	-2.45 (0.54) ***	-0.23 (0.11)
Condition × Multi-tasking	0.60 (0.22)**	0.70 (0.26)**	-4.69 (4.04)	-1.09 (0.58)	0.05 (0.10)
Non-verbal intelligence	-1.42 (1.23)	-2.54 (1.37)	-71.52 (21.68)**	-4.39 (2.44)	-0.14 (0.52)
Counterbalance order	0.25 (0.43)	0.71 (0.48)	4.80 (7.54)	-0.97 (0.84)	-0.07 (0.18)
Experience	-0.05 (0.74)	0.40 (0.83)	-12.98 (12.99)	-1.99 (0.58)	-0.40 (0.31)

Note: *** $p < .001$, ** $p < .01$, * $p < .05$. The estimates represent standardised coefficient estimates for fixed effects in the linear mixed model on the trial level specifying a random intercept and slope per participant. P-values were computed using the Wald approximation.

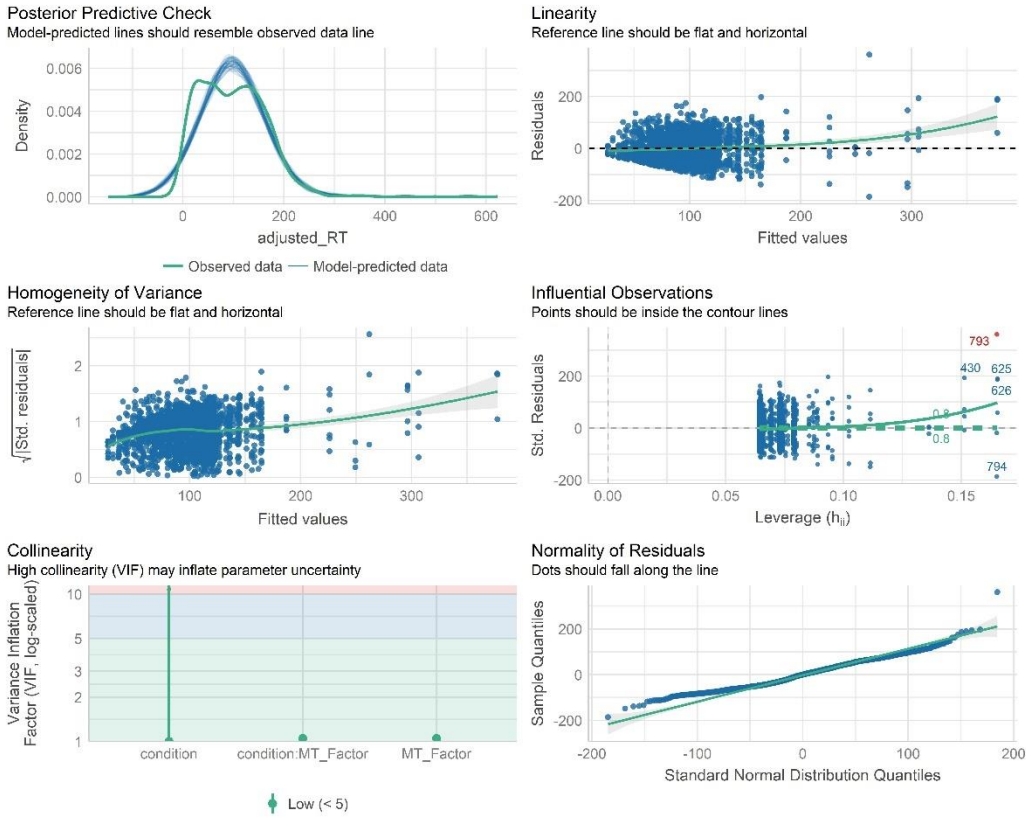
Assumption plots: Acceptances



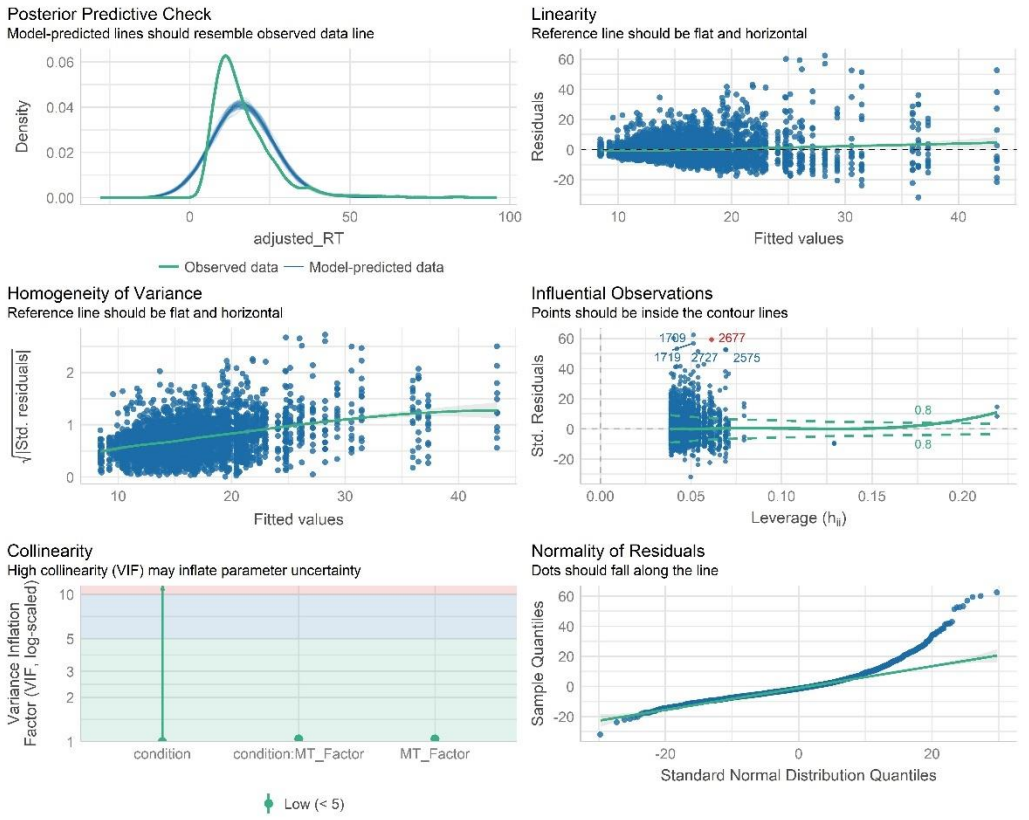
Hand-offs



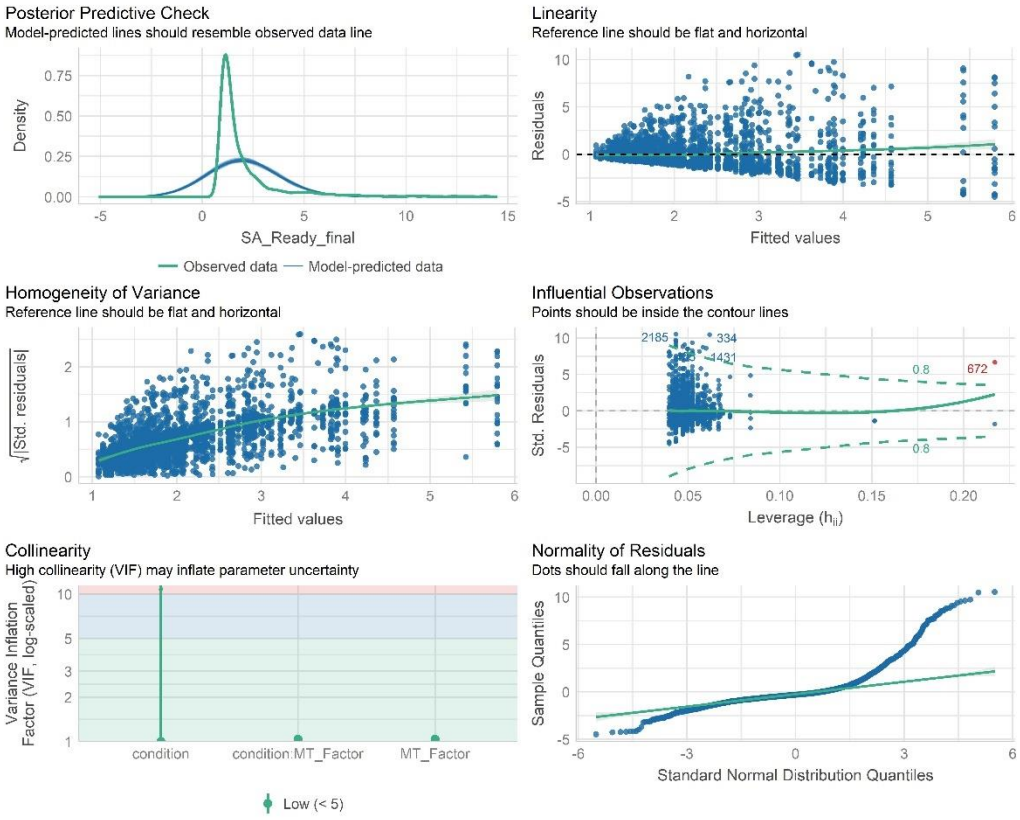
Conflict detection



Situation Awareness



SA Ready – Objective workload



Chapter 3:

Here is presented the SA queries used in Chapter 2 and Chapter 3. These tables include the results of the analyses discussed in Chapter 3 for comparison.

On page 95 I note that the Visuospatial Tracking Task was included, along with the previously described cognitive tasks. The written instructions participants received before commencing the task are presented here.

Visuospatial Tracking Task:

Task Instructions

This task examines how good you are at multitasking – doing two tasks simultaneously.

Thresholding Phase

During the first phase which lasts 12 minutes, you will practice tracking a large dot with the mouse for 30 second trials. The dot will turn green when your mouse on the dot and red when it is not to provide feedback on your performance. At the end of each trial you will be given your % accuracy to track your performance.

You will also complete a Go/ No Go task in which you are shown a target shape and colour, then a series of shapes will flash rapidly on the screen, and you will have to press space bar when the target shape appears. The area around the shape will turn green when you correctly press space to identify the target, and when you do NOT press space when the target is not present. The area will turn red when you incorrectly press space bar when the target is NOT present or when you don't press space bar when the target is present. These trials last 60 second. At the end of each trial you will be given your % accuracy to track your performance.

You will complete 4 of each of these trials. The program will track your performance. You should aim for 80% accuracy in all trials. The program will increase the level of difficulty if you are significantly above 80% and decrease the difficulty if you are significantly below 80%.

At the end of the thresholding, the level of difficulty at which you achieved 80% will be used in the test phase.

Testing Phase

During the testing phase all trials last 60 seconds. You will complete the Tracking Task and the Go/No Go task as you did in the thresholding phase, but without any feedback (no green for correct trials/ on target tracking).

You will also get a third type of trial in which the tracking and Go/ No Go are to be completed at the same time. It is your job to give equal attention to both tasks (i.e. Do NOT just focus on tracking and ignore the shapes and visa versa).

The test phase will last 21 minutes

Try your best to stay at 80% accuracy in all trials. The trials with both will be hard, but that is to test your multitasking ability!

Table 4: SA queries Scenario 1, including percent of participants who correctly answered the queries (accuracy), and scale Crombach's Alpha, Table is color coded Green for queries that scored highly on the analyses and are considered 'good', red for queries that scored poorly on the analyses and are considered 'poor'. Uncolored queries are neither good or poor. Grey query was removed for incorrect coding.

Queries		**Accuracy**	**Scale Alpha if item deleted**
1	Are aircraft EK23 and aircraft AA31 currently located in the NW, NE, SW or SE quadrant	0.96	0.747
2	What aircraft is on the same flight level as aircraft SQ57	0.85	0.751
3	In which quadrant will the next loss of separation take place within the next 30 secs if no action will be taken?	0.60	0.779
4	What is the flight level of the aircraft that you last accepted in to the sector in the NE quadrant?	0.87	0.737
5	What aircraft is on the same flight level as aircraft AA37 in the SW quadrant?	0.86	0.754
6	Will aircraft SQ79 and aircraft VA32 cross path in the NW, NE, SW, or SE quadrant?	0.94	0.742
7	How many aircraft needed accepting within the last 30 secs in the SW quadrant?	0.65	0.771
8	Which quadrant currently has the most traffic/number of aircraft?	0.94	0.746
9	What is the flight level of the two aircraft that you last accepted in to the sector?	0.95	0.747
10	Closest to which waypoint will aircraft AA83 and aircraft NZ55 cross path?	0.97	0.755
11	What is the next waypoint that QF40 has to cross?	0.93	0.742
12	What is the flight level of the aircraft that you last accepted in to the sector?	0.94	0.75

13	Will aircraft QF59 and aircraft EK85 cross path in the NW, NE, SW or SE quadrant?	0.96	0.743
14	Are aircraft VA28 and SQ56 travelling at the same speed?	0.82	0.764
15	Which quadrant currently has the most traffic/number of aircraft?	0.27	
16	How many aircraft did need handing off within the last 30 seconds (28,20,58)?	0.64	0.765
17	Will aircraft VA95 and aircraft NZ33 cross path in the NW, NE, SW or SE quadrant?	0.95	0.749
18	What common waypoint will aircraft SQ51 and aircraft QF81 both pass?	0.94	0.743

Table 5: SA queries Scenario 3, including percent of participants who correctly answered the queries (accuracy), Crombach's Alpha, Inter-Item correlations and Structure Loading in Item Cluster Analysis. Table is color coded Green for queries that scored highly on the analyses and are considered 'good', red for queries that scored poorly on the analyses and are considered 'poor'. Uncolored queries are neither good or poor.

Queries		Accuracy	Scale Alpha if item deleted
1	What aircraft is on the same flight level as aircraft QF45 in the NE quadrant?	0.91	0.763
2	In which quadrant will the next loss of separation take place if no action as been or will be taken w4ithin the next 30 secs	0.72	0.762
3	What is the flight level of the aircraft that you last accepted to the VA route.	0.68	0.749
4	What common waypoint will aircraft QF94 and aircraft AA36 both pass?	0.95	0.751
5	At what altitude is aircaft NZ20 travelling	0.95	0.751
6	What is the flight level of the two aircraft that you last acepted to the QF and SQ route?	0.92	0.741
7	**What waypoint are aircraft QF59 and SQ27 currently closest to?**	0.93	0.729
8	Will aircraft AA58 and aircraft NZ17 cross path in the NW, NE, SW, or SE quadrant?	0.97	0.74
9	What waypoint do aircraft AA35 and QF52 both have to cross?	0.97	0.752
10	Are aircraft SQ47 and VA37 travelling at the same speed?	0.81	0.777

11	What is the next waypoint that QF87 has to cross?	0.97	0.754
12	What is the flight level of the aircraft that you last accepted on to the EK route?	0.87	0.744
13	What is the flight level of the aircraft that you last accepted to the VA route?	0.93	0.76
14	Are aircraft AA63 and aircraft NZ19 currently located in the NW, NE, SW or SE quadrant?	0.94	0.737
15	Which quadrant currently has the most number of aircraft inside the flight sector?	0.95	0.757
16	In which quadrant will the next loss of separation take place if no action as been or will be taken within the next 30 secs	0.73	0.758
17	What waypoint does the last accepted aircraft on the AA route have to cross next (AA42)	0.79	0.744
18	What aircraft did you last hand off on the QF route?	0.92	0.742

Table 6: New SPAM Situation Awareness Queries used in Chapter 3, counterbalanced across type and task.

Scenario 1		
Type	Task	Question
Present	Conflict	Q1 - Are aircraft EK23 and aircraft AA31 currently located in the NW, NE, SW or SE quadrant?
Past	A/H	Q2 - What was the last waypoint SQ57 had to cross?
Present	A/H	Q3 - Is aircraft QF62 currently located in the NW, NE, SW, or SE quadrant?
Past	Conflict	Q4 - What was the flight level of the aircraft that you last accepted in to the sector in the NE quadrant?
Future	A/H	Q5 - How many aircraft will need handing off in the next 30 seconds in the SE quadrant?
Future	A/H	Q6 - Will aircraft SQ79 and aircraft VA32 cross path in the NW, NE, SW, or SE quadrant?
Past	A/H	Q7 - How many aircraft have you accepted into the SW quadrant in the last 60 seconds?
Present	A/H	Q8 - Which quadrant currently has the most number of aircraft in the flight sector?
Past	Conflict	Q9 - What were the flight level of the two aircraft that you last accepted in to the sector?
Future	Conflict	Q10 - What common waypoint will aircraft AA83 and aircraft NZ55 cross path closest to?

		Q11 - What was the last waypoint that QF40 has to cross?
Past	A/H	Q12 - What was the flight level of the aircraft that you last accepted in to the
Past	Conflict	NE quadrant?'
		Q13 - Will aircraft QF59 and aircraft EK85 cross path in the NW, NE, SW
Future	Conflict	or SE quadrant?
		Q14 - Are aircraft VA28 and SQ56 travelling at the same speed?
Present	Conflict	Q15 - How many aircraft will need handing off in the next 30 seconds?
Future	A/H	Q16 - what common waypoint will aircraft E88 and E99 have to pass?
		Q17 - Will aircraft VA95 and aircraft NZ33 cross path in the NW, NE, SW
Future	Conflict	or SE quadrant?
		Q18 - What common waypoint will aircraft SQ51 and aircraft QF81 both
Future	Conflict	pass?
Future	A/H	

Scenario 3		
Past	A/H	Q1 – How many aircraft have been accepted in the last 60 seconds?
Future	Conflict	Q2 Will aircraft QF73 and VA53 cross paths in the NW, NE, SW or SE quadrant?
Past	Conflict	Q3 - What is the flight level of the aircraft that you last accepted to the QF route?
Future	Conflict	Q4 - What common waypoint will aircraft QF94 and aircraft AA36 both pass?
Present	A/H	Q5 - At what flight level is aircraft NZ20 travelling?
Past	A/H	Q6 - What was the flight level of the two aircraft that you last accepted to the QF and SQ route?
Present	Conflict	Q7 - What waypoint are aircraft QF59 and SQ27 currently closest to?
Future	Conflict	Q8 - Will aircraft AA58 and aircraft NZ17 cross path in the NW, NE, SW, or SE quadrant?
Future	A/H	Q9 - What waypoint do aircraft AA35 and QF52 both have to cross?
Past	Conflict	Q10 – What was the flight level of the last two aircraft you accepted in the SW quadrant?
Future	A/H	Q11 - What is the next waypoint that QF87 has to cross?
Past	Conflict	Q12 - What was the flight level of the aircraft that you last accepted to the EK route?
Past	Conflict	Q13 - What was the flight level of the aircraft that you last accepted to the VA route?
Present	A/H	

Future	A/H	Q14 - Are aircraft AA63 and aircraft NZ19 currently located in the NW, NE, SW or SE quadrant?
Present	Conflict	Q15 - How many aircraft will need handing off in the next 30 seconds?
Future	A/H	Q16 - Is aircraft AA79 currently located in the NW, NE, SW and SE quadrant?
Past	A/H	Q17 - What waypoint does the last accepted aircraft on the AA route have to cross next?
		Q18 - What aircraft did you last hand off on the QF route?

In chapter 3 (p. 100-101) I describe the linear mixed models used to test the hypotheses. In the first instance, model comparisons are presented here show the comparison between a null model and the interaction model for each dependent variable. This demonstrates that models with the interaction of condition and multi-tasking were the models with the greatest predictive power.

Null model formula: dependent variable ~ (1 | participant_number) + (0 + condition | participant_number)

*Interaction model formula: dependent variable ~ condition * MT_Factor + (1 | participant_number) + (0 + condition | participant_number)*

Table 7: Model comparisons between null model and full model for each dependent variable model using Chi-square test of significance.

Model	AIC	BIC	Log Likelihood	Deviance	Chisq	Df	Significance
Acceptances Null	43360	43389	-21676	43352			
Acceptances Model Full	43203	43253	-21594	43189	163.60	3	<.001
Hand-offs Null	47278	47307	-23635	47270			
Hand-offs Model Full	47224	47274	-23605	47210	60.62	3	<.001
Conflict Null	23257	23268	-11625	23249			
Conflict Model Full	23229	23268	-11607	23215	34.92	3	<.001
SA Null	24409	24433	-12200	24401			
SA Model Full	24387	24430	-12186	24373	28.41	3	<.001

Objective Workload Null	12882	12906	-6436.90	12874			
Objective workload Full	12885	12928	-6435.6	12871	2.68	2	0.44

Note: Log likelihood is a measure of fit describing how well a parameter explains the observed data. AIC is Akaike Information Criterion, an extimator of prediction error. Lower AIC reflects better model fit. BIC is Bayesian Information Criterion, a measure for model comparison or selection. Lower BIC reflect better model fit. Chisquare is a test of significant difference between models.

Additionally, models were conducted which included previous ATC experience, condition order, non-verbal intelligence and subjective rating of task importance were included as covariates.

*Formula: Dependent variable ~ condition * MT_Factor + IQ + Counterbalance + Experience + (1 | participant_number) + (0 + condition | participant_number)*

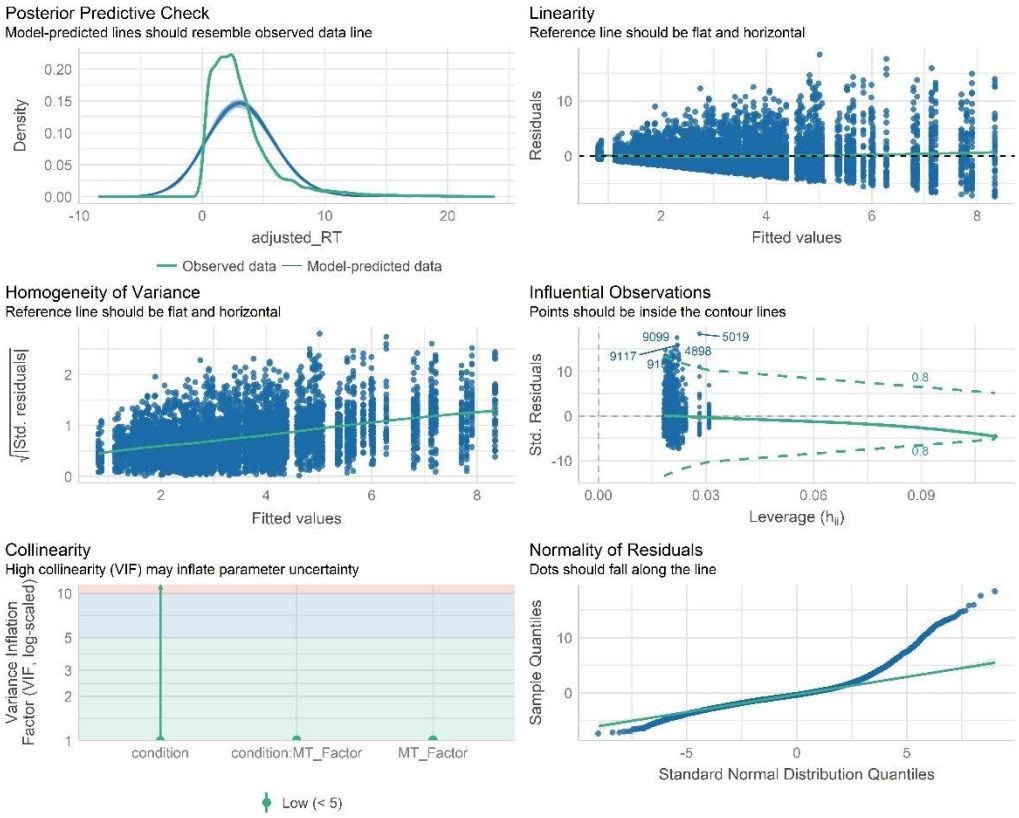
These were included to control for potential confounding effects. These covariate models are presented here. None of these covariates influenced the results of the main effects of interest.

Table 8. Covariate models unstandardised regression effects. Standard deviation presented in parentheses.

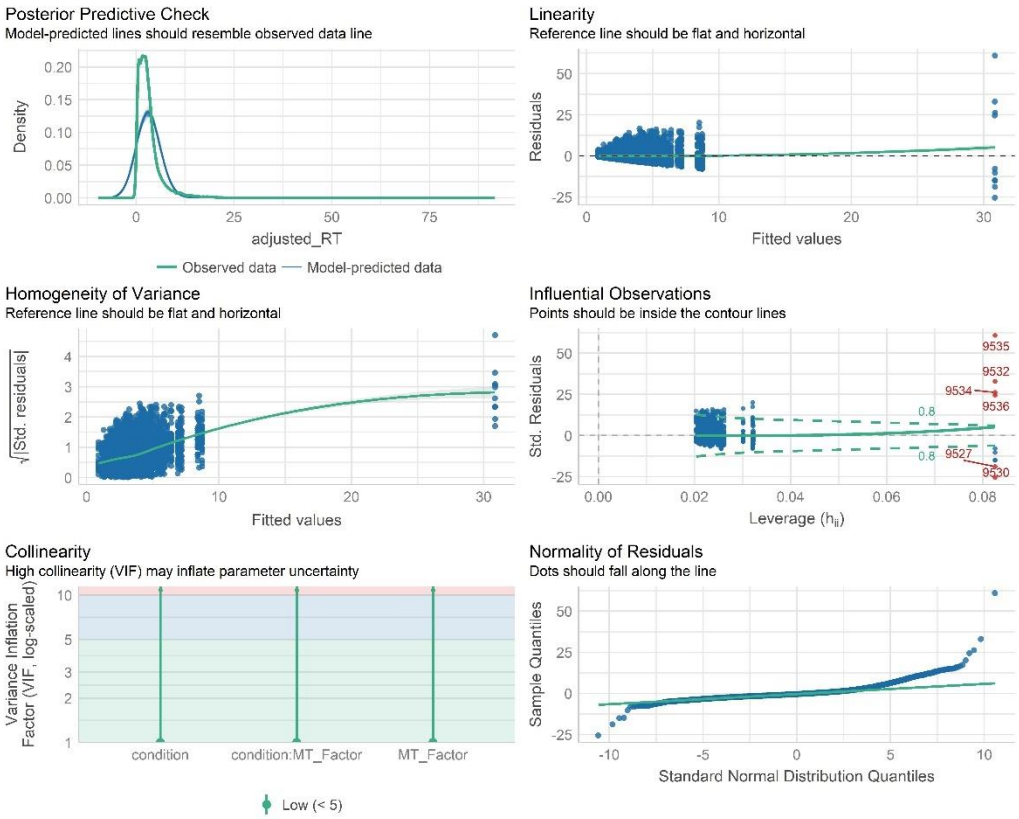
Parameter	Acceptances Adjusted RT	Hand-offs Adjusted RT	Conflict Adjusted RT	SA Question Adjusted RT	Objective Workload
Condition	-1.81 (0.10) ***	-1.38 (0.24) ***	-18.16 (2.98) ***	-0.47 (0.35)	-0.01 (0.07)
Multi-tasking	-0.61 (0.14) ***	-1.39 (0.30) ***	-4.61 (3.80)	-3.00 (0.53) ***	-0.12 (0.08)
Condition × Multi-tasking	0.37 (0.12) **	1.17 (0.28) ***	4.92 (3.55)	0.26 (0.41)	0.06 (0.08)
Non-verbal intelligence	-1.51 (0.51) **	-2.48 (1.09) *	-18.70 (12.96)	-1.24 (1.85)	-0.51 (0.27)
Counterbalance order	-0.47 (0.23)	-0.84 (0.51)	1.67 (6.07)	-0.35 (0.87)	0.00 (0.13)
Experience	-0.25 (0.31)	-0.34 (0.67)	-15.04 (8.00)	-3.33 (1.14) **	-0.30 (0.17)

Note: *** $p < .001$, ** $p < .01$, * $p < .05$. The estimates represent standardised coefficient estimates for fixed effects in the linear mixed model on the trial level specifying a random intercept and slope per participant. P-values were computed using the Wald approximation.

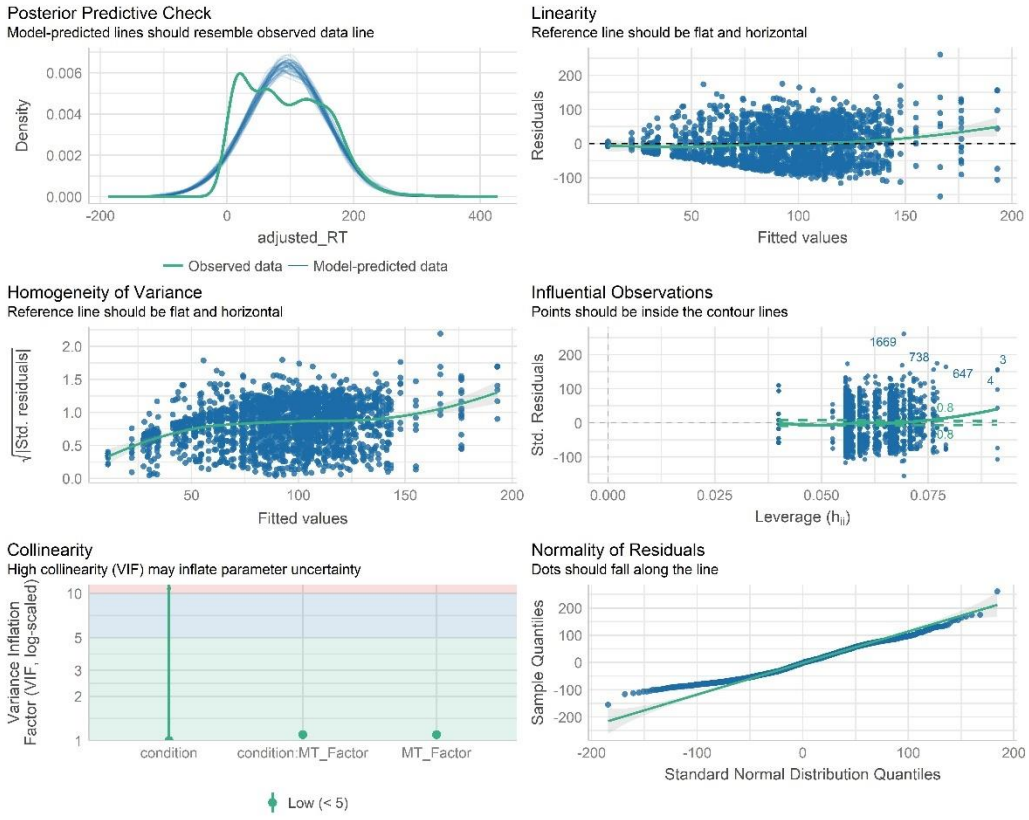
Assumption plots: Acceptances



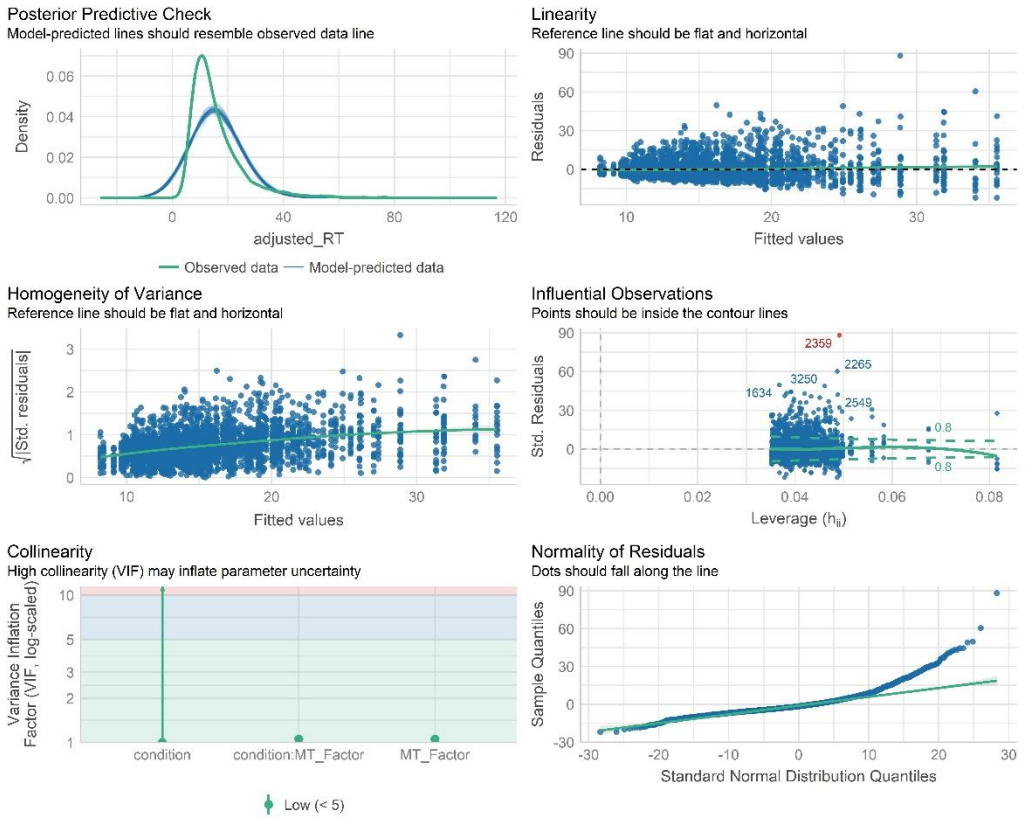
Hand-offs



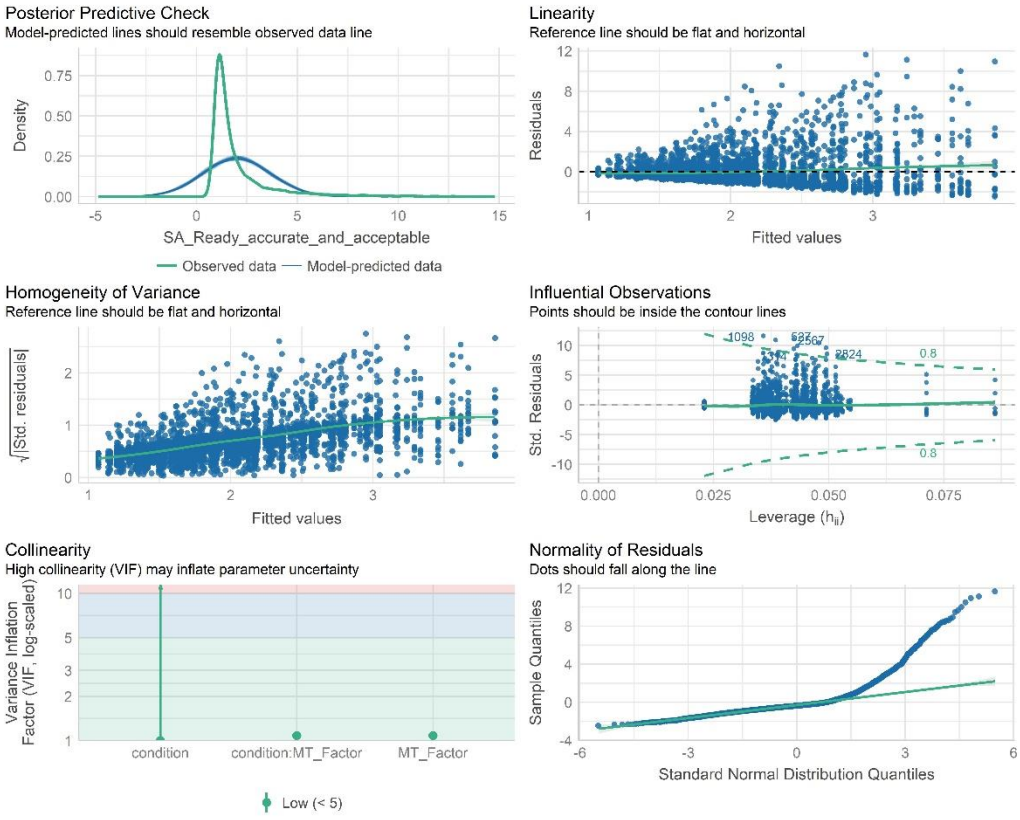
Conflict Detection



Situation Awareness



SA Ready – Objective workload



Chapter 4

In chapter 4 (p. 130) I describe the changes to the ATC task which included the addition of automation errors. Below is a timeseries list of when errors occurred in relation to other events in the task.

Table 9: Count of events in ATC (at 30 second intervals), including number of Failure Events

Scenario 1

Time (mins)	Acceptances	Hand-offs	Conflicts	Failures
0.5	1			
1	2	1		
1.5		1		
2		1		
2.5	3	1		
3	1	1	1	
3.5		1		
4	1			
4.5	1	1		
5	1	1		
5.5	1			1A
6		1		
6.5	1			
7	2	2	1	1H
7.5				
8	1	1		
8.5		1		
9	1	2		
9.5				
10	1		1	
10.5	1	1		
11	1	1		1A
11.5	1			
12		2		1H
12.5	1	1		
13	1		1	
13.5	2			1A

14				
14.5	1	2	1	1H
15	2			
15.5				
16		2		1H
16.5	1	1		1A
17		1	1	
17.5				
18	2			1A
18.5	1	3		1H
19	1	1		
19.5		1		
20	1	1		1H
20.5	1	1		1H
21	1	1		1A
21.5	1	1	1	1H
22				
22.5	1	1		1A
23		1		
23.5			1	
24	1	1		1A
24.5		1		
25	1	1		1H
25.5				
26			1	
26.5	2	2		1A/ 1H
27				
27.5	1	1		1H
28	1	1		1A
28.5			1	
29				
29.5	1	2		1A/ 1H
30				
30.5				
TOTAL	44	46		14 ACCEPTANCES/ 14 HANDOFFS

Note: A = acceptance failure event, H = hand-off failure event

Scenario 3

Time (mins)	Acceptances	Hand-offs	Conflicts	Failures
0.5		1		
1	1	1		
1.5	1			
2		1		
2.5	2	1		
3	2	1		
3.5	1	1		
4			1	
4.5		1		
5	1	1		
5.5	1			
6				
6.5		2	1	
7	1	1		1A
7.5	1			
8		1		
8.5	1	1	1	
9	1	1		
9.5	1			
10	1	2		1H
10.5	2	1		1A
11		1		
11.5	1		1	
12		1		
12.5	1	1		1A
13	1			
13.5	1			
14	1	1		1H
14.5		2		1H
15	1	1	1	
15.5				
16	1	1		1A

16.5	1	1		1H
17	1	1	1	1A
17.5				
18	1			
18.5	1	2		1H
19	1			
19.5	1			
20		1	1	
20.5	1	1		1A
21	1			
21.5	1	2		1H
22		1	1	1H
22.5	1			
23	2	1		1A
23.5				
24		1		1H
24.5	1	1		1A
25				
25.5	1	3		2H
26	2	1	1	1A
26.5				
27		1		
27.5	2			1A
28	1			1A
28.5		1	1	1H
29		2		1H
29.5				
30		1		1H
TOTAL	43	46	10	14 ACCEPTANCES/ 14 HANDOFFS

In chapter 4 I describe the linear mixed models used to test the hypotheses. In the first instance, model comparisons are presented here show the comparison between a null model and the interaction model for each dependent variable. This demonstrates that models with the interaction of condition and multi-tasking were the models with the greatest predictive power.

Null model formula: dependent variable ~ (1 | participant_number) + (0 + condition | participant_number)

*Interaction model formula: dependent variable ~ condition * MT_Factor + (1 | participant_number) + (0 + condition | participant_number)*

*Interaction model formula with trial type variable: dependent variable ~ condition * MT_Factor + trial_type * MT_Factor + (1 | participant_number) + (0 + condition | participant_number)*

Table 10: Model comparisons between null model and full model for each dependent variable model using Chi-square test of significance.

Model	AIC	BIC	Log Likelihood	Deviance	Chisq	Df	Significance P
Conflict Null	24548	24526	-12236	24540			
Conflict Model Full	24486	24526	-12236	24472	68.12	3	<.001
SA Null	24663	24687	-12327	24655			
SA Model Full	24643	24687	-12315	24629	25.23	3	<.001
Objective Workload Null	13973	13997	-6982.3	13965			
Objective workload Full	13974	14017	-6980.0	13960	4.69	3	.19
Acceptances Null	49852	49881	-24822	49844			
Acceptances Full Model with Trial Type	49580	49630	-24783	49566	278.61	3	<.001
Hand-offs Null	50553	50582	-25273	50545			
Hand-offs Full Model with Trial Type	49863	49914	-24925	49849	695.72	3	<.001

Note: Log likelihood is a measure of fit describing how well a parameter explains the observed data. AIC is Akaike Information Criterion, an extimator of prediction error. Lower AIC reflects better model fit. BIC is Bayesian Information Criterion, a measure for model comparison or selection. Lower BIC reflect better model fit. Chisquare is a test of significant difference between models.

Additionally, models were conducted which included previous ATC experience, condition order, non-verbal intelligence and subjective rating of task importance were included as covariates.

*Formula: Dependent variable ~ condition * MT_Factor + IQ + Counterbalance + Experience + (1 | participant_number) + (0 + condition | participant_number)*

These were included to control for potential confounding effects. These covariate models are presented here. None of these covariates influenced the results of the main effects of interest.

Table 11: Covariate models unstandardised regression effects. Standard deviation presented in parentheses.

Parameter	Conflict Adjusted RT	SA Question Adjusted RT	Objective Workload
Condition	-31.17 (3.80) ***	-1.23 (0.36) ***	0.00 (0.08)
Multi-tasking	-12.84 (4.56) ***	-1.73 (0.49) ***	-0.19 (0.12)
Condition × Multi-tasking	7.72 (4.87)	0.47 (0.45)	0.02 (0.10)
Non-verbal intelligence	8.82 (17.79)	-3.40 (1.96)	-0.42 (0.47)
Counterbalance order	-2.79 (6.54)	-0.46 (0.72)	0.00 (0.17)
Experience	-4.30 (13.18)	0.09 (1.48)	-0.46 (0.36)

Note: *** $p < .001$, ** $p < .01$, * $p < .05$. The estimates represent standardised coefficient estimates for fixed effects in the linear mixed model on the trial level specifying a random intercept and slope per participant. P-values were computed using the Wald approximation.

In Chapter 4 additional analyses were conducted with the trial type (manual, reliable automation or failure automation) added as a covariate. Models with the previously described covariates are presented here.

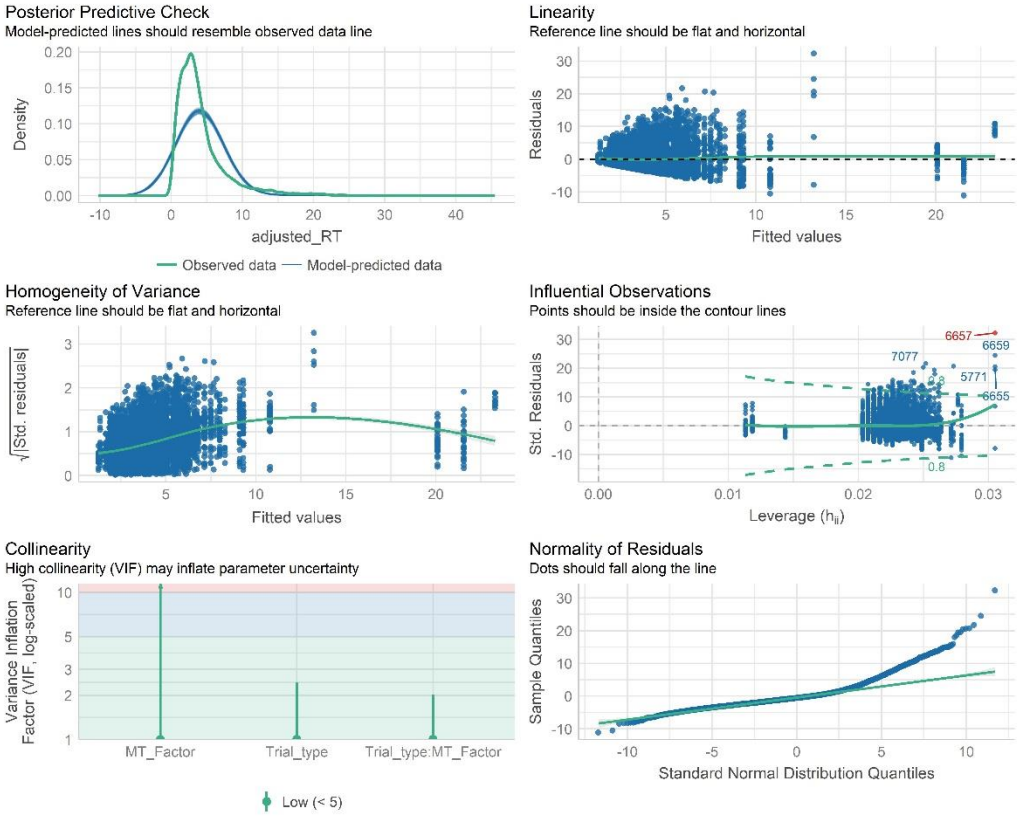
*Formula: Dependent ~ condition * MT_Factor + Trial_type + IQ + Counterbalance + experience + (1 | participant_number) + (0 + condition | participant_number).*

Table 12: Covariate models unstandardised regression effects with trial covariate for acceptances and hand-offs. Standard deviation presented in parentheses.

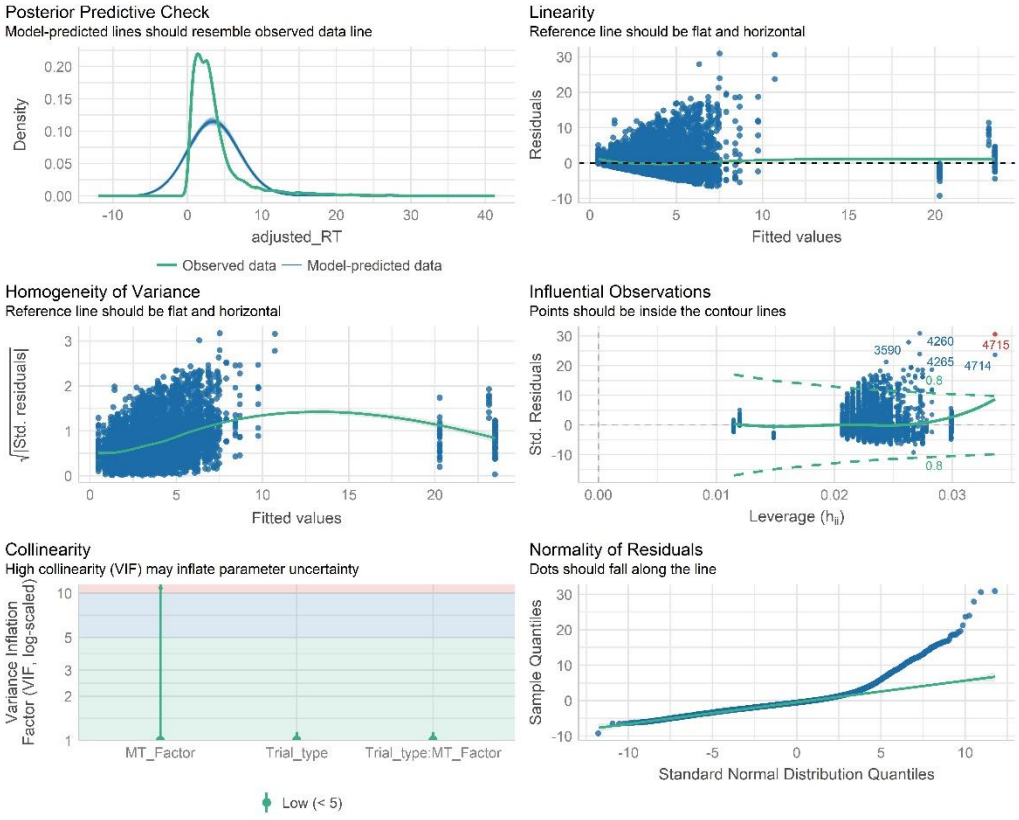
Parameter	Acceptances Adjusted RT	Hand-offs Adjusted RT
Condition	-3.57 (0.16) ***	-4.64 (0.16) ***
Multi-tasking	-0.43 (0.25)	-0.60 (0.27) *
Condition × Multi-tasking	0.81 (0.20) ***	1.35 (0.21) ***
Trial Type	2.16 (0.09) ***	3.19 (0.10) ***
Trial Type × Multi-tasking	-0.67 (0.12) ***	-0.92 (0.13) ***
Non-verbal intelligence	0.18 (1.03)	-0.36 (1.09)
Counterbalance order	-0.32 (0.38)	-0.29 (0.40)
Experience	2.40 (0.20) **	3.08 (0.81) ***

Note: *** $p < .001$, ** $p < .01$, * $p < .05$. The estimates represent standardised coefficient estimates for fixed effects in the linear mixed model on the trial level specifying a random intercept and slope per participant. P-values were computed using the Wald approximation.

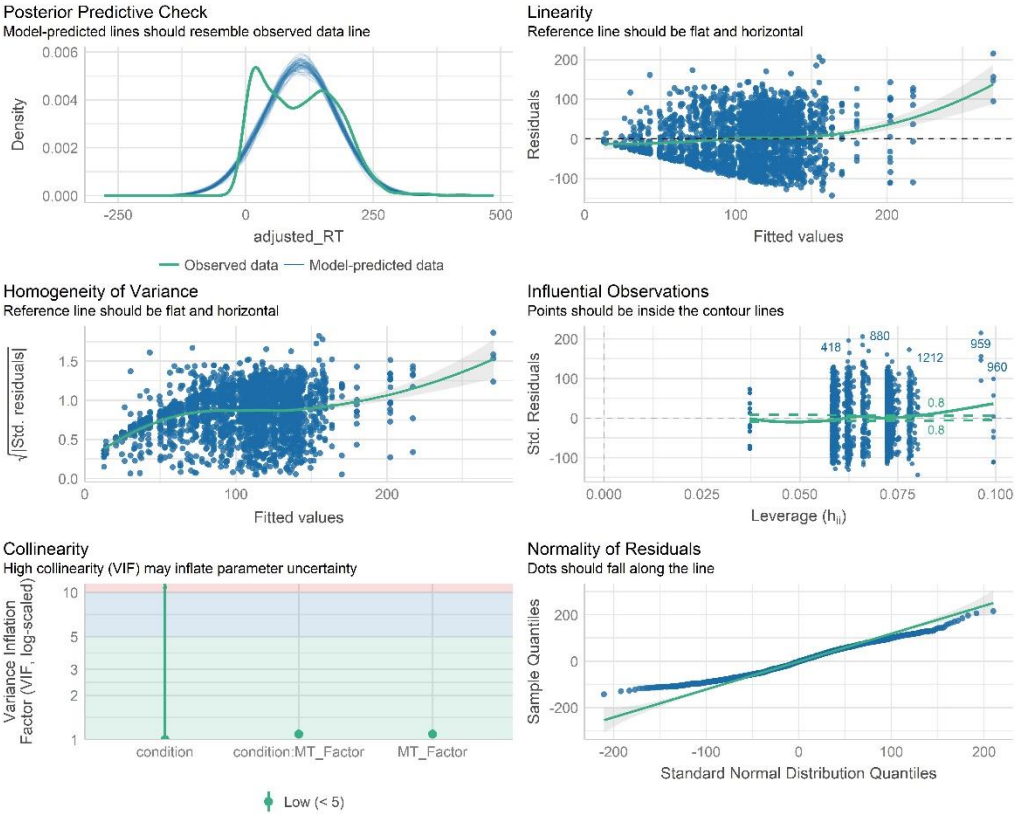
Assumption Plots: Acceptances by trial



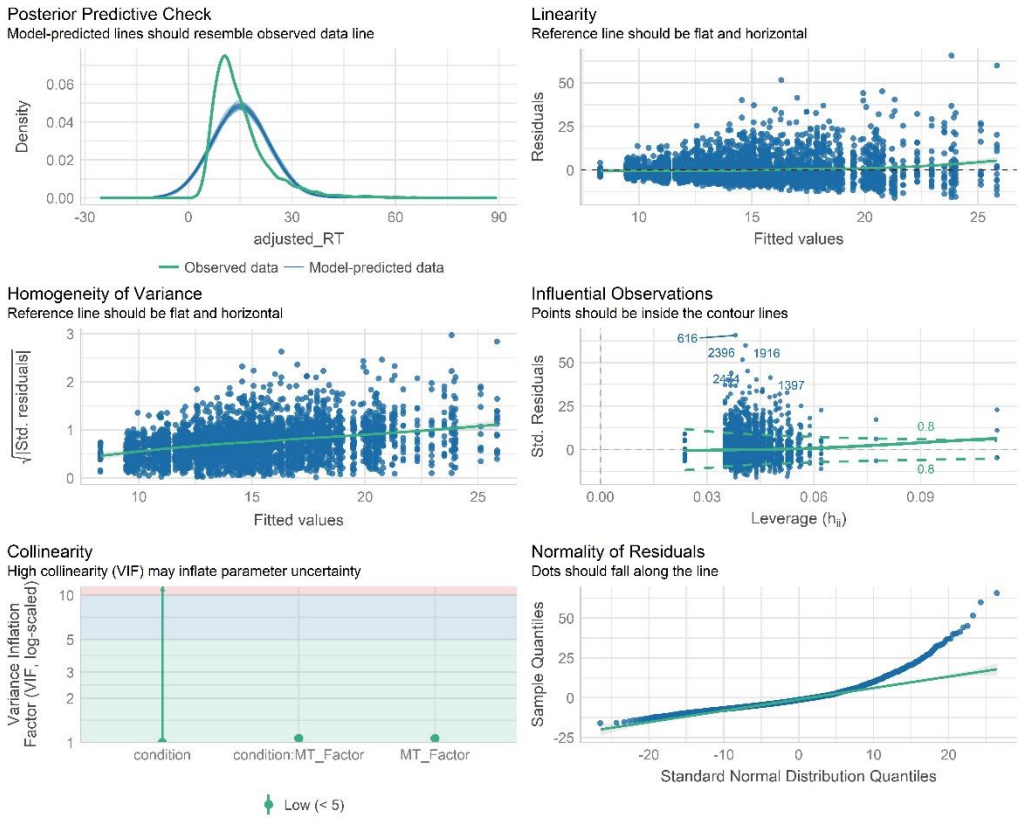
Hand-offs by trial



Conflict detection



Situation Awareness



SA Ready – Objective workload