Behavioural Measurement of Cognitive Workload

by

Reilly James Innes, BPsyc (Hons I) (*Newcastle*)

Supervised by:

Prof. Scott D. Brown; A Prof. Ami Eidels;

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy(Psychology)

This research was supported by an Australian Government Research Training Program (RTP) Scholarship

March 1, 2021

Statements

- *Originality:* I hereby certify that the work embodied in the thesis is my own work, conducted under normal supervision. The thesis contains no material which has been accepted, or is being examined, for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made. I give consent to the final version of my thesis being made available worldwide when deposited in the University's Digital Repository, subject to the provisions of the Copyright Act 1968 and any approved embargo.
- *Authorship:* I hereby certify that the work embodied in this thesis contains published paper/s/scholarly work of which I am a joint author. I have included as part of the thesis a written declaration endorsed in writing by my supervisor, attesting to my contribution to the joint publication/s/scholarly work.

Signed: Reilly Innes

Date: March 1, 2021

By signing below I confirm that Reilly Innes contributed [insert description / outline of contribution] to the paper/ publication entitled [insert reference details]

Signed: Scott Brown

Date: March 1, 2021

Acknowledgements

I referred to Song, Kang, Timakum, and Zhang (2020) for guidance on common acknowledgement practice and have thus limited my acknowledgements to follow the key patterns, and average length, as identified by the CNN+Doc2Vec algorithm, as well as relying on common example keywords (shown in *italics*).

1. Peer interactive communication and Technical support

First and foremost, I would like to thank my supervisor Professor Scott Brown and co-supervisor Associate Professor Ami Eidels for their continued *support* and belief in me throughout my honours and PhD candidacy. Further, I am thankful for all of the *academic assistance, fruitful discussions, valuable suggestions, insightful comments and resources* that you have provided. I will always see you as the ultimate academic role models.

I would also *like to acknowledge other academics* who in some way, shape or form contributed to this thesis – Dr. Z. Howard, Dr. N. Evans, Mr. A. Thorpe, Ms. C. Kuhne, Mr. G. Cooper, Dr. K. Nesbitt, Dr. L. Wall, Dr. P. Garrett, Dr. G. Hawkins, Mr. J-P. Cavallaro, Ms. G. Newcombe and Ms. J. Sparre; without your assistance this would have taken a lot longer.

Additionally, I wish to acknowledge the members of the Newcastle Cognition Lab – *thank you for providing* a space for which I truly felt at home in my work.

2. Financial

This research was supported by an Australian Government Research Training Program (RTP) Scholarship awarded to RI, for which he is very *grateful*.

3. General acknowledgement

Finally, I would like to acknowledge my loved ones. To my parents – *thank you for supporting me*, *encouraging* my curiosity, *teaching* me the important things in life and having a seemingly unlimited workload capacity. To my brother – thank you for always being the C to my R (literally). To Sarah – thank you for putting up with me and filling my heart as full as my workload in a four dots to track MOT paradigm. To my friends – thank you for providing the necessary rest breaks between blocks and feedback I needed. Thank you all for your support, which feels like one of my experiments to a participant – endless.

List of Publications

The work within this thesis has lead to the following journal articles that are either currently published, submitted, or in preparation, which I have listed with the full bibliographic citations in the order they appear in the thesis:

1. Howard, Z., Innes, R., Brown, S. D., & Eidels, A. (2018). Cognitive workload and analysis of flight path data. *Technical Report*

2. Innes, R., Howard, Z., Eidels, A., & Brown, S. D. (2018). Cognitive workload measurement and analysis. *Technical Report*

3. Innes, R. J., Howard, Z. L., Evans, N. J., Eidels, A., & Brown, S. D. (*2020*). A broader application of the detection response task to cognitive tasks and online environments. *Human Factors*. doi: `https://doi.org/10.1177/0018720820936800`

4. Innes, R. J., Howard, Z. L., Thorpe, A., Eidels, A., & Brown, S. D. (*2020*). The effects of increased visual information on cognitive workload in a helicopter simulator. *Human Factors*. doi: `https://doi.org/10.1177/0018720820945409`

5. Innes, R. J., & Kuhne, C. L. (2020). An LBA account of decisions in the multiple object tracking task. The Quantitative Methods for Psychology, 16, 175–191. doi: `10.20982/tqmp.16.2.p175`

Statement of Contribution

Below I have included a statement outlining both my contribution, and the involvement of others, for each chapter where the research performed involved collaboration. This addresses the requirements of both the *Statement of Collaboration* and the *Statement of Authorship*. The below statement have been endorsed by my primary supervisor, Professor Scott Brown.

Publication Contributions

Chapter 3

Innes, R. J., Howard, Z. L., Thorpe, A., Eidels, A., & Brown, S. D. (*2020*). The effects of increased visual information on cognitive workload in a helicopter simulator. *Human Factors*. doi: `https://doi.org/10.1177/0018720820945409`

- *Reilly Innes:* Planned design (25%), conducted study (35%), analysed results (45%), wrote paper (70%).
- *Zachary Howard:* Planned design (25%), conducted study (35%), analysed results (45%), wrote paper (20%), edited paper (20%).
- *Alexander Thorpe:* Conducted study (30%), analysed results (10%), wrote paper (10%), edited paper (10%).
- *Ami Eidels:* Planned design (25%), edited paper (40%).
- *Scott Brown:* Planned design (25%), edited paper (30%).

Chapter Contributions

- *Chapter 3:* I was involved in the development of the Experiments outlined in Chapter 3. In conjunction with Scott Brown, Ami Eidels and Nathan Evans, I designed the original DRT-MOT experiment (as part of another project which is now published – see Innes, Evans, Howard, Eidels, and Brown (2020)), which was extended in Chapter 3. Nathan Evans assisted with initial programming of the design. I completed all of the testing and analysis for Experiment 1. Gemma Newcombe and Jessica Sparre assisted with data collection and analysis for one of the subsequent experiments in Chapter 3. I programmed and analysed both of the subsequent experiments of Chapter 3, most of which is not included here for brevity, but can be found at `https://osf.io/ayp6d/` .

- *Chapter 5:* This project was work in collaboration with Airbus Helicopters and Hensoldt Sensor Systems. Airbus personnel proposed the initial problem (helicopter pilot overload). In collaboration with Zachary Howard, Scott Brown and Ami Eidels, I assisted in refining the research question and methodology. I also completed the literature review. Alexander Thorpe, Keith Nesbitt, Ami Ediels, Scott Brown, Zachary Howard and myself completed the data collection. Zachary Howard completed the majority of flight data analysis, whilst Alexander Thorpe also contributed. I completed the majority of DRT analyses. I was also responsible for producing a technical report of cognitive workload measurement (see Innes, Howard, Eidels, and Brown (2018)), and contributed to the technical report led by Zachary Howard for the flight data analysis (see Howard, Innes, Brown, and Eidels (2018)). I was involved as the lead author for the resulting publication (Innes, Howard, Thorpe, Eidels, & Brown, 2020), with comments from Zachary Howard, Ami Eidels, Alexander Thorpe and Scott Brown, whom are also co-authors on the publication. This chapter is published at *Human Factors: The journal or human factors and ergonomics society.*

- *Chapter 6:* I was involved in establishing the working relationship with the ADF group involved in testing, which has continued over the four years of my candidature. This project also had contributions from Zachary Howard, Ami Ediels and Scott Brown. I developed the experiment and proposed and conducted the analysis, and this was assisted by Ami Eidels and Zachary Howard. I collected data from the undergraduate

participants and was responsible for identifying similar and/or useful literature for this chapter.

- *Chapter 7:* I completed the literature review, analysis and writing for chapter 7. Caroline Kuhne assisted with model analysis and was a co-author of a resulting paper (see Innes and Kuhne (2020)). The paper is published at *The Quantitative Methods for Psychology Journal*. I conducted the joint model analysis using insights from this paper. Scott Brown provided comments and direction for the analysis and discussion. Gavin Cooper assisted with the sampling process.

Additional Publications

Listed are additional journal articles or internal reports that are either currently published, submitted, or in preparation, which I have been involved in during my candidature. These publications are listed here as they are closely related to the work shown in this thesis, however were part of different projects and/or theses.

1. Howard, Z. L., Evans, N. J., Innes, R. J., Brown, S., & Eidels, A. (2019, August 27). How is multitasking different from increased difficulty? *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-020-01741-8

2. Thorpe, A., Innes, R. J., Townsend, J., Heath, R., Nesbitt, K., & Eidels, A. (2020) Assessing Cross-Modal Interference in the Detection Response Task. *Journal of Mathematical Psychology, 98, 102390.*

Additional Work

Listed are additional publications and presentations which have relevance to the thesis, but are not included in it, or represent earlier iterations of included work. In each case, the presenter's name is written in **bold**:

1. **Innes, R. J.**, Eidels, A., & Brown, S. (2017). *Evidence for an Applied Measure of Cognitive Capacity in a Laboratory Environment.* Experimental Psychology Conference. Shoal Bay, NSW, AUS.

2. **Innes, R. J.**, Howard, Z. L., Eidels, A., & Brown, S. (2017). *An objective measure of cognitive workload: Evaluation and practical Application.* CBMHR Postgraduate Conference. Newcastle, NSW, AUS.

3. **Innes, R. J.**, Howard, Z. L., Eidels, A., & Brown, S. (2018). *Biting off more than you can process: Group differences in cognitive workload.* Experimental Psychology Conference. Hobart, TAS, AUS.

4. **Innes, R. J.**, Howard, Z. L., Thorpe, A., Eidels, A., & Brown, S. (2019). *Flying blind: Does adding information really help?* Australasian Mathematical Psychology Conference. Melbourne, VIC, AUS.

5. **Innes, R. J.**, Howard, Z. L., Eidels, A., & Brown, S. (2019). *A measurement tool for comparing cognitive workload differences* (poster). Society for Computers in Psychology Conference. Montreal, CAN.

6. **Innes, R. J.**, Kuhne, C. L., & Brown, S. (2020). *Modelling decisions of the multiple object tracking task.* Australasian Mathematical Psychology Conference. Coogee, NSW, AUS.

7. **Innes, R. J.**, & Brown, S. (2020). *Joint modelling group differences from military personnel.* 2020 Annual Meeting of the Society for Mathematical Psychology. Virtual Conference.

8. **Cooper, G.**, Cavallaro, J., Innes, R. J., Kuhne, C., Hawkins, G., & Brown, S. (2020). *Hierarchical Bayesian parameter estimation with the Particle Metropolis within Gibbs sampler.* 2020 Annual Meeting of the Society for Mathematical Psychology. Virtual Conference.

Contents

Abstract

Everyday we are faced with a myriad of tasks to complete, ranging from the most simple to highly complex. Our ability to complete such tasks is limited by inherent mechanisms which allow us to focus our attention and perceive the optimal amount of information needed to do so. As more information becomes available, and as a result of our propensity to multitask, these cognitive limits are pushed and stretched. In doing so, we often ignore important task relevant information, or our performance is inhibited. To fully understand the interplay of these factors, we need to be able to measure and evaluate workload. In this thesis I investigate the construct of cognitive workload, which is inherently limited by our overall capacity, through a measure used predominantly in applied driver distraction literature. From this, I present a body of work that expands upon theoretical underpinnings and new applications of this measure. In the theoretical stream, I show the usefulness, reliability, and applicability, of this measure in lab-based scenarios, whilst in the applied stream, I show three novel uses of the measure in both theoretical and real-world scenarios, as well as developing analyses applicable to such scenarios. The research in this thesis has implications and applications across a broad range of research areas, ranging from theoretical, in areas such as methodological development, to highly applied, in areas such as aviation environment evaluation.

In the interest of openness and replicability, all data (from student cohorts)[1], analysis and further appendices from this thesis can be found at `https://osf.io/ayp6d/`.

[1]RAAF group data (Chapter 6 and 7) is confidential and cannot be shared publicly. The same applies to data collected from Airbus Helicopters & Hensoldt Sensor Systems (Chapter 5).

"There are a great many people in the country today, who through no fault of their own, are sane. Some of them were born sane, while others became sane later in their lives. It is up to people like you and me, who are out of our tiny little minds, to help them overcome their sanity."

– The Reverend Arthur Belling

Chapter 1

Defining Cognitive Workload & Capacity

Multitasking has become the norm, where we constantly find ourselves juggling a number of task demands at once. We have an inherent belief in our own ability to multitask, and further we assume that we are *more* efficient when doing more things at once. However, research has shown that this heuristic is far from true, suggesting that when multitasking, we are actually rapidly switching our attention between tasks, meaning we increase our margins of error on *both* tasks (Adler & Benbunan-Fich, 2012; Pashler, 2000). Our margin for error may decrease as tasks are practiced and become more automatic, or the error margin may increase as tasks become more complex or more mental demand is required. Measuring this margin for error in multitasking is the crux of this thesis.

When performing difficult tasks, greater concentration and mental effort is required to attend to key information (Gevins, Smith, McEvoy, & Yu, 1997). Similarly, if you perform multiple simple tasks, increased mental effort is required to switch between tasks effectively and avoid missing key events (Gopher, Armony, & Greenshpan, 2000; Meyer & Kieras, 1997; Monsell, 2003). This "mental effort" is often termed *cognitive workload* (Lee, Young, & Regan, 2008). Cognitive workload is of significant importance when undertaking any task, yet is often overlooked or ignored by researchers, designers and policy makers. Cognitive workload is important to performance as we are limited in the amount of information and tasks we can attend to, so overloading ourselves with task demands leads to increased error.

We are inherently limited in our capacity to perceive information and operate on perceptual input. This limited capacity is often referred to as our *cognitive capacity* (Eidels, Donkin, Brown, & Heathcote, 2010; Kahneman, 1973). Cognitive scientists have for many years attempted to explore the limit of human perception and attention, and there are a myriad of studies and theories exploring conceptual underpinnings of attention, such as the limits of our perception (Palmer, 1990; T. D. Wickens, 2002), how this perceptual limitation varies from our attentional capacity (A. Treisman & Geffen, 1967), the bottleneck of information between perception and attention (Pashler, 1984) and a variety of contributing factors. These factors have been extensively studied within the field of attention, however, in this thesis, the key assumption which I rely upon is that cognitive capacity is, in some way, limited.

Furthermore, we are limited in how we *respond* to incoming information (Meyer & Kieras, 1997). Although modern day environments allow for greater multitasking behaviour,

our ability to respond is similarly inherently limited (Pashler, 1994). One main stream of multitasking literature has indicated that when completing multiple tasks, we do not purely "multitask" or perform two tasks concurrently, but rather rapidly switch our attention between tasks (Gopher et al., 2000). Another theory of attention posits the idea of "threaded cognition" (Salvucci & Taatgen, 2008), where streams of thought for each concurrent task are "threaded" together by a serial procedure and then executed across available resource channels. From the task switching account, when task switching occurs, it is apparent that key information may be missed, or responses may not be effectively executed. This is the second key tenet of this thesis – not only is cognitive capacity inherently limited, but our ability to respond to more than one task is also inherently limited (Pashler, 1998). Furthermore, these factors share key elements; the limits of capacity affect the limits of responding.

These limiting factors, and the nature of interaction between scenarios, individuals and contexts can affect on-task performance. Take for example a driving environment. The driving environment already requires a great amount of attentional control and may use the majority of one's cognitive capacity. Maintaining speed, monitoring for hazards and constantly adjusting position are just some of the tasks occupying a driver's attention, with events occurring frequently and at high speeds. If a driver was to also text on their mobile phone, or engage in a phone conversation, then task performance could be affected (Strayer et al., 2013; Strayer & Johnston, 2001). Primarily, in perceiving information from the phone, they may be limiting the information perceived from the driving environment. Furthermore, by rapidly switching between the phone and driving, potentially critical responses, such as braking to an unexpected hazard, or information, such as perceiving a mindless pedestrian, could be easily missed (Coleman, Turrill, Cooper, & Strayer, 2016; Strayer, Cooper, Turrill, Coleman, & Hopman, 2017; Strayer, Drews, & Johnston, 2003). Driver distraction research shares many links with situational awareness literature. Situational awareness relates to perception of the state of an environment and the elements and events within that environment (Gilson, 1995), and so cognitive workload is a key moderator of situational awareness – as conditions of high workload may limit situational awareness (Selcon, Taylor, & Koritsas, 1991; Tsang & Vidulich, 2006), but low workload and monotony can also be detrimental to situational awareness (Hancock & Matthews, 2019). Evidently, balancing workload demands is key to optimising performance and situational awareness. In the context of situational

awareness literature, it is clear that understanding workload is essential to safety. Further, in environments such as aviation or driving, cognitive workload has a key role in situational awareness – an essential construct for safe operation (C. D. Wickens, 2002a). In order to understand these effects, a valid and reliable measure of cognitive workload is a necessity.

The overarching structure of this thesis is as follows, with early chapters devoted to validating a design to test a measure of cognitive workload, and later chapters focusing on extending that design to answer new theoretical, and practical, research questions. For convenience to the reader, I limit my following literature review to key concepts reviewed throughout the thesis, whilst going into further detail on concepts specifically related to each chapter within that chapter. The following section aims to define the key concepts used throughout the thesis and form a thread which links together all of the individual components.

1.1 Defining Key Concepts

Cognitive capacity refers to the overall amount of attentional resources available to an individual (Townsend & Wenger, 2004b). Any action taken by an individual imposes some load on overall capacity (Pashler, 2000). The load imposed on cognitive capacity is known as *cognitive workload* (C. D. Wickens, 2008). The amount of cognitive workload experienced may vary between individuals (Hart, 2006; Jaeggi et al., 2007) and is affected by a variety of factors, however these can be generalised to task difficulty or the number of concurrent tasks (C. D. Wickens, 2002a). Task difficulty can encompass the complexity of the task, the amount of mental effort required to complete it or even the amount of distraction the individual faces. The number of tasks simply relates to how many activities the individual is engaged in. With these factors in mind, consider that cognitive workload is inherently restricted by cognitive capacity (Kahneman, 1973). If one's cognitive workload exceeds their cognitive capacity, a scenario of "overload" is experienced (Jaeggi et al., 2007; C. D. Wickens, 2008). Furthermore, in scenarios of extremely low cognitive workload, an individual may be subject to cognitive "underload" (Lavie, 2010). Both scenarios are detrimental to task performance – however, cognitive overload is seemingly more detrimental, and consequently is the focus of the current research, as well as literature in situational awareness (Stanton, Chambers, &

Piggott, 2001; C. D. Wickens, 2002a), driving environments (Strayer et al., 2013; Strayer & Johnston, 2001), autonomous driving environments (Biondi et al., 2018; McKerral, Boyce, & Pammer, 2019), aviation (Berka et al., 2007; Svensson, Angelborg-Thanderez, Sjöberg, & Olsson, 1997; Wilson, 2002), user-interface (Gross, Bretschneider-Hagemes, Stefan, & Rissler, 2018; Thorpe, Nesbitt, & Eidels, 2019), military (Huttunen, Keränen, Väyrynen, Pääkkönen, & Leino, 2011) and others.

1.2 Cognitive Capacity

The mental limitation when completing tasks is often referred to as our *cognitive capacity* (Eidels, Townsend, Hughes, & Perry, 2015; Townsend & Eidels, 2011). Cognitive capacity is large enough so that we can process a useful amount of incoming information from our environment, but restricted as to limit unnecessary stimuli (Eidels et al., 2010). The consequence of limited cognitive capacity is that if a resource channel amasses more information than it can handle, it will become overloaded. Generally, this is known as a cognitive bottleneck – where only some information survives in mental processes following an attentional filter (Borst, Taatgen, & Van Rijn, 2010; Salvucci & Taatgen, 2008). Subsequent performance of a task in the presence of overload will be negatively impacted according to Kahneman (1973) due to a loss of task and information awareness. Cognitive capacity limitations have been central in almost all major theories of attention (see Broadbent, 2013; Kahneman, 1973, for examples), perception (see A. M. Treisman & Gelade, 1980, for examples), and memory (see Baddeley & Hitch, 1974; Miller, 1956, for examples).

1.3 Cognitive Workload

In this thesis, I define Cognitive workload as the cognitive demand faced by an individual at any single time. Cognitive workload is limited by capacity and defined by the number and difficulty of the tasks being done at any one time (Kahneman, 1973). It may also encompass and be affected by the interference between tasks (in a multitasking paradigm) and distractors from the environment (De Jong, 2010). For example, an easy task with no distractions would elicit a low level of cognitive workload, whereas someone undertaking

an easy task in a distracting environment, would likely experience a high cognitive workload. Similarly, when completing more difficult tasks or a greater number of tasks, cognitive workload increases.

There are many common examples throughout diverse literature fields where multitasking, increased difficulty and greater distractions have an impact on cognitive workload. Strayer et al. (2013) showed that adding the operation span task increased cognitive workload in a driving task, similar to the effects of talking to extra passengers. Similarly, Strayer et al. (2017) showed the distracting effects of in-vehicle entertainment systems. In educational research, Brünken, Steinbacher, Plass, and Leutner (2002) and Brünken, Plass, and Leutner (2004) showed the workload inducing effects of multi-modal learning stimuli, and in user-interface literature, the nature of displays can negatively affect workload (Brock, Stroup, & Ballas, 2002; Gross et al., 2018; Thorpe et al., 2019). We see similar effects on cognitive workload when the task difficulty is increased rather than the number of tasks increased (Engström, Larsson, & Larsson, 2013). For example, increasing the difficulty of a n-back task showed similar effects to increasing the number of tasks on cognitive workload (Mehler, Reimer, Coughlin, & Dusek, 2009; Strayer, Watson, & Drews, 2011; Young, Hsieh, & Seaman, 2013). These examples highlight both the sensitive nature of cognitive workload and the range of environments where workload changes may be detrimental to performance.

Similar to high cognitive workload, low cognitive workload can also have detrimental effects on performance (Hancock & Matthews, 2019). In scenarios where the tasks are monotonous or when engagement is low, performance may suffer. This notion forms the opposite end of the inverse U-shaped arousal curve, where performance is low when workload is too high or too low, and so balancing task demands here is critical in optimising performance and avoiding errors.

Errors due to high cognitive workload can be hugely detrimental in some environments. Take for example driving. There is a vast amount of research that highlights the negative outcomes of using a mobile phone in the driving environment (Engström, Åberg, Johansson, & Hammarbäck, 2005; Strayer et al., 2013, 2003; Strayer & Johnston, 2001). The importance of research investigating driver distraction has been exemplified in policy changes over the preceding decades, with new laws and regulations shaping the driving environment (Strayer et al., 2011; Young et al., 2013). As technology continues to develop, it is essential

to consider the impact of technology on cognitive workload, as technology can significantly contribute to errors. This research shows prime examples of multitasking behaviour leading to erroneous and detrimental task performance, and highlights the importance of validating workload measures. Further, as Feigh, Dorneich, and Hayes (2012) outline in their framework for developing adaptive interfaces, measuring workload plays a key role in identifying operator performance decrements in order to create intelligent adaptive interfaces. These adaptive processes could involve limiting or increasing the level of information available to operators, moderating task schedules to aid performance or implementing automated processes to assist with the task.

It is evident that cognitive workload is a vital consideration in driving environments, and there are many more areas where workload impacts performance, such as aviation and education. In fact, any field within human factors research, it could be theorized, is subject to cognitive workload factors (Thorpe et al., 2019). For example, in aviation, pilots complete routine flight checklists before take off and could be easily distracted by extra tasks or environmental factors. In this environment, cognitive workload needs to be adequately controlled, and capacity trained, to allow personnel to deal with adverse circumstances. Further studies on cognitive workload have shown the importance of workload in areas such as forklift operation (Gross et al., 2018), helicopter piloting (Gaetan et al., 2015) and air traffic control (Ahlstrom & Friedman-Berg, 2006; Marek, Karwowski, & Rice, 2010). The majority of cognitive workload research has tended to focus on distracted driving, and situational awareness literature. Many of these studies have shown how dual-task cognitive workload measures can be applied to new environments, with consistent reliability. However, so far there have been minimal studies conducted in controlled, in–laboratory settings, without complex paradigms. Typically, this scenario is reversed – where we take well researched, reliable and valid designs from the lab and apply them to real world environments. A common issue arising from this research workflow is that cognitive theories are developed in isolation from real world contexts. Consequently, when applying these models and theories to real world environments, we find that they often lack tractability. Here, I take a paradigm that is valid and reliable in real world settings and has simple, quantifiable, underlying theory, and use this to develop and test cognitive models and theories which otherwise could not be advanced in the field. This includes adapting such theory for new purposes and across previously unexplored domains, as well as developing models which capture holistic performance.

Evidently, there is great value in developing theory in this fashion, as we are more aware of the validity and applicability of such measures whilst the underpinning theory remains simple and tractable. This thesis focuses on dual-task performance as a central (though not standalone) method of cognitive workload assessment and aims to develop theory from here. The following section details the dual-task methodology which will be used throughout this thesis.

1.4 Dual-Tasks

Dual task paradigms allow researchers to measure cognitive workload from primary tasks which lack clear, quantifiable outcome measures. Generally the additional task is structured in such a way to not interfere with the main task, and instead maximise the performance metrics which reflect primary task cognitive workload. These measures reflect cognitive workload given the limited capacity of attention which restricts performance across simultaneous tasks.

In the dual-task method of measuring cognitive workload (which will be discussed in more detail in Chapter 2), participants complete a main task – such as driving – and a secondary task, which measures the amount of residual capacity is available (Conti, Dlugosch, Vilimek, Keinath, & Bengler, 2012; C. D. Wickens, 2002a). Performance on *both* tasks provides an indication of the participants cognitive workload, as performance in the main task cannot be sacrificed to perform well in the secondary task. That is, performance in the secondary task can measure cognitive workload, provided that performance in the primary task is maintained. Detriment to performance in the secondary task may indicate that the primary task is highly cognitively loading, and therefore requiring a greater amount of attentional resources available.

It is also possible that two tasks may require simultaneous (or highly frequent) responding, consequently leading to responses affected by the psychological refractory period, rather than workload (Pashler, 1994). The psychological refractory period is the period of time immediately following a response, where subsequent responses are inhibited momentarily. Tasks that require more frequent responding are often perceived as being more difficult,

and consequently dual-task measures of workload may be compromised by the psychological refractory period. In the following chapter (see Chapter 2), I show how dual-task methodology has overcome such difficulty, using continuous tasks which require constant responding, or through using primary tasks which increase workload without requiring simultaneous responses.

The detection response task (DRT) is the task most commonly used in dual-task distracted driving paradigms (Castro, Strayer, Matzke, & Heathcote, 2019; Innes, Evans, et al., 2020), where participants are asked to detect and respond to a salient signal (the full method of this design will be discussed in Chapter 2). Results from driving studies (for examples see Strayer et al. (2013), Engström et al. (2005) and Merat and Jamson (2008)), show an increase in DRT response times with increasing conditions of difficulty. Increased DRT response times have also been associated with increased distraction such as using mobile phones (Strayer et al., 2013), using in vehicle information systems (Coleman et al., 2016) and communicating with smart assistants (Strayer et al., 2019). The DRT methodology follows the dual-task method, measuring the residual attentional capacity from the main task/s in an attempt to quantify cognitive workload induced by said task/s. Despite the frequent responses required by the DRT, there is no evidence that responding is significantly affected by the psychological refractory period (Thorpe et al., 2019), and further, studies employing the DRT often use a continuous secondary task (Engström et al., 2013; Stojmenova & Sodnik, 2018), as advocated above.

1.5 Thesis Experiment General Framework

For all experiments in this thesis, I use the DRT methodology in dual-task scenarios. Further, to overcome the possible interference caused by the psychological refractory period, I use primary tasks which are free of simultaneous responding (or in the case of Chapter 5, require continuous responding). Throughout this thesis, I will use one experimental paradigm for all chapters, with the exclusion of Chapter 5. The DRT-MOT methodology was taken from Innes, Evans, et al. (2020) (also see Howard, Evans, Innes, Brown, & Eidels, 2020, for a similar example). Only changes to the Innes, Evans, et al. (2020) Experiment 2 methodology will be outlined in subsequent method sections.

1.6 Thesis Overview

Figure 1.1 provides a visual overview of the structure of this thesis. It is well known that human capacity is limited (Kahneman, 1973), yet in the modern era, we are pressed to process more information and complete more tasks, more efficiently. Evidence shows that multitasking comes with an associated cost, often to performance, yet still we continue to push against, or even ignore our limits (Kahneman, 1973; Pashler, 1994). Despite our adaptability, and ability to rapidly switch between tasks effectively, it is evident that multitasking still incurs a cost (Strayer et al., 2013). Evaluating this cost is crucial to the efficiency and safety of task completion.

To address this need, one focus of the theoretical portion of this thesis is to extend upon pre-existing measures of cognitive workload. This methodology extension can be seen in the top third of Figure 1.1. In Chapter 2, I outline a number of cognitive workload measures which have been proposed and tested. Importantly, the measures tend to agree on the general workload inducing factors and the magnitude of workload increase. From here, I outline a dual-task methodology of cognitive workload which I use throughout this thesis. Chapter 3 provides validation for an in-lab, dual-task method using the DRT (to measure cognitive workload) and the multiple object tracking task (MOT; to induce workload); a paradigm I refer to as the "DRT-MOT". In this chapter I show evidence for: validity and reliability – across varying DRT signal modalities; construct validity – in comparing to a validated measure; and external reliability – by comparing to a divergent construct measure. In essence, Chapters 2 and 3 form a theoretical platform for the thesis, where I establish a rationale, reliability and validity for the paradigm.

Further, I aim to expand the scope of the DRT beyond distraction evaluation. This notion can be seen in the middle portion of Figure 1.1, where the research splits into two streams. In Chapter 4, I highlight the usability of this measure in a lab-based setting and, in Chapter 5, I show practical applications of this methodology. These chapters show how environmental factors can affect cognitive workload, and provide an evaluative methodology. Chapter 4 evaluates the effects of two types of assistance on cognitive workload and in Chapter 5, using a similar methodology, I evaluate heads-up display information in a helicopter simulator. This experiment provides a practical application of DRT methodology

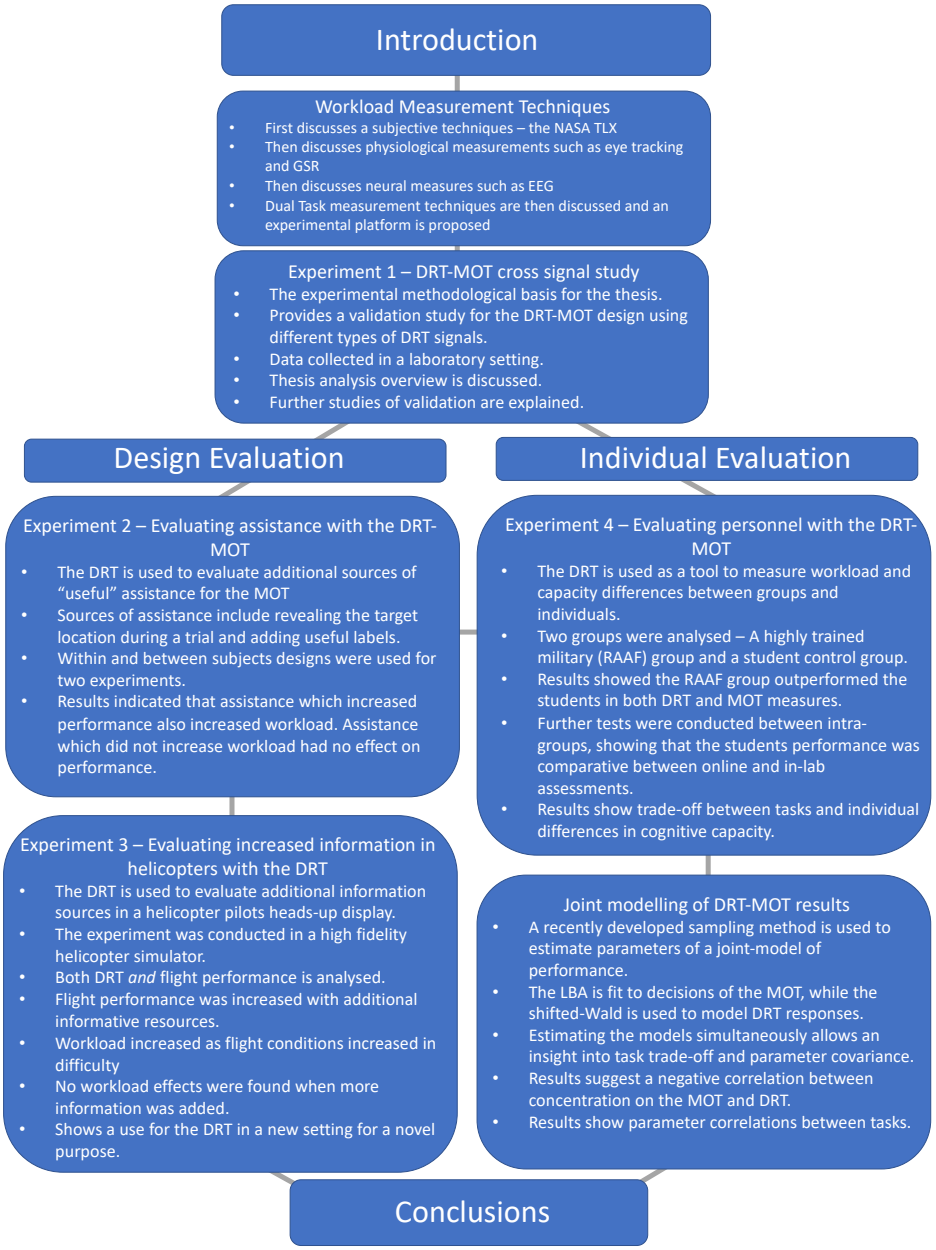


FIGURE 1.1: Overview of this thesis. There is a common experimental thread followed throughout (Experiment 1), which is used for several novel purposes. The thesis splits into two main components, one component shows the usefulness of the DRT for new evaluative purposes (left side), whilst the other component uses the DRT to differentiate between people and groups (right side). This differentiating component seeks to not only use the measure for a novel purpose, but to extend our understanding of human processing. This is somewhat extrapolated in the lower portion of the right pathway, where data from Experiment 4 is fit using computational cognitive modelling techniques. The streams were performed simultaneously throughout the duration of my candidature. This forms a clear overall sequence of research which seeks to broaden the scope of cognitive workload literature and dual-task methodology.

to an alternate context and to evaluate "usefulness" rather than distraction. These studies show complimentary results in that one type of information can prove useful to the participant, but often at the cost of workload; whilst another type of information may provide no performance boost at no workload cost. These chapters establish a useful extension of cognitive workload measurement, by going beyond distraction evaluation.

Chapter 6 shows a further application of dual-task cognitive workload measurement. In this chapter, I had privileged access to a highly trained group of military personnel who completed the DRT-MOT paradigm during a selection period for a sought after role. Rather than only evaluating differences between the military group and a control undergraduate student group, I use results to distinguish between individuals. In distinguishing between individuals, not only is the sensitivity of the paradigm shown, but overall cognitive capacity can be inferred in a novel way. Additionally, validity of the DRT-MOT paradigm is furthered, with results showing no differences between online and in-lab participants, as well as establishing a benchmark for the military group. This study shows a further use of the DRT methodology – to infer cognitive capacity as a selection – and provides greater insight into the cognitive underpinnings of individual performance.

Chapter 7 then uses modelling techniques to extrapolate individual strategy differences that may underpin results of Chapter 6. Furthermore, using new modelling techniques, I provide a joint model analysis of the DRT-MOT task. Through the joint model framework, not only do I show differences between conditions of load, but I also show individual differences and correlations *between processes* across tasks. These results allow a deeper understanding of the the cognitive processes that underpin performance differences between groups and individuals. Using this joint modelling technique is a novel approach to analysis from cognitive workload tasks, and results emphasise the usefulness of such analysis.

Together, Chapters 1, 2 and 3 form a theoretical stream in a combined review and validation for existing measures of cognitive workload. Chapters 4, 5, 6 and 7 highlight new uses of cognitive workload measurement tools – as shown in the middle third of Figure 1.1 – where novel applications and analysis techniques allow for a greater understanding of the impact of cognitive workload and underpinning latent cognitive processes. This methodology stream emphasises the usefulness of cognitive workload evaluation for applications further

to those currently observed and provides measurement and analysis tools within a dual-task framework for a range of purposes and contexts. This stream represents an important contribution to the literature in both extending the role of cognitive workload evaluation and in understanding cognitive workload effects in new environments. Finally, Chapter 8 provides general conclusions concerning the main aims and issues discussed throughout the thesis.

Chapter 2

Measuring Cognitive Workload & Capacity

There have been a wide variety of attempts to measure and quantify cognitive workload during both laboratory (i.e. theoretical) and applied tasks. As outlined in Chapter 1, cognitive workload intuitively has a large impact on a range of domains, so measurement and quantification are vital to improve systems, task performance, safety protocols and policy (Gawron, 2019). These investigated measures generally have associated positives and negatives, which will be elucidated within this chapter. It should be noted though that of the wide range of research domains there is relatively little convergent evidence of appropriate measures of workload. The most commonly used method appears to be subjective measures of workload based on self-report (Hart, 2006). Research has continually shown the limitations of subjective methods of experimentation, so it is surprising that alternate cognitive workload measures are relatively under-explored (Paulhus, 1991). Other, more objective, measures of workload involve evaluating performance, neural activity and psycho-physiological activity.

There are evidently a range of cognitive workload measures, which have all shown some sensitivity to workload fluctuations, as outlined by Matthews, Reinerman-Jones, Barber, and Abich IV (2015), however, the convergent validity between these measures is poor. A further point made in this paper is the criteria for a valid cognitive workload measure, which should be considered when proposing new measures, or applying existing measures, as divergent workload measures have strengths and weaknesses across contexts. For a full review of the criteria of a valid workload measure, see (Matthews et al., 2015). As will be outlined in the following sections, each measure of cognitive workload has benefits and drawbacks, which may determine the usability of the measure for given contexts. One major drawback of many workload measures is that measures of cognitive workload resulting from a primary task may fail to account for main task performance. Generally we could focus on primary task performance as an indicator of workload, however, in many applied scenarios where we wish to assess workload, the primary task is without quantifiable or clear outcome measures. Secondly, a general limitation of all measures of cognitive workload is that there is high between-subject variability. This is the case in many psychological research areas, and is overcome by using within-subjects designs. This form of experimental methodology is equally valid across all types of workload measures, and should be considered when reading this chapter. There are few, if any, cognitive workload studies that rely on between-subjects designs and comparisons.

The following chapter provides a brief overview of measures of cognitive workload which have been used previously. These include the subjective questionnaire method of the NASA task load index (TLX), neural measures such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI), bio-metric measures such as galvanic skin response (GSR) and heart rate, and conclude with dual task measures – including the detection response task (DRT).

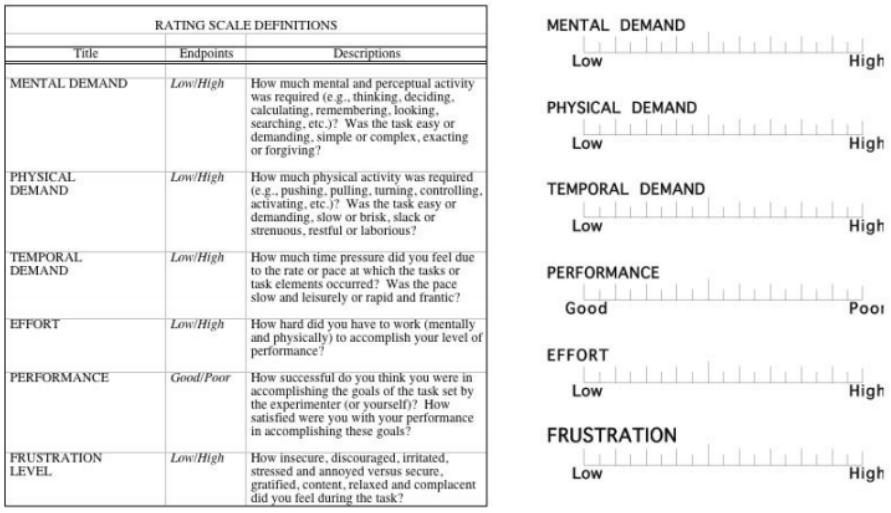
2.1 Subjective Measures

Subjective measures of psychological phenomena have been widely used for decades, with surveys and questionnaires offering introspective insights into behaviour and reasoning. In cognitive workload measurement, subjective measures of workload are highly common and have dominated the literature since the development of the NASA Task Load Index (TLX) (Hart, 2006). There are several subjective methods of cognitive workload that have been used throughout the literature, often to measure the impact of a task on cognitive workload. Loft et al. (2018) however, used the Air Traffic Workload Input Technique to predict upcoming workload demands – a method which has potential for adaptive interface design. The most common subject measure of workload however, remains the NASA TLX, which I outline in the section below.

2.1.1 NASA Task Load Index

The NASA task load index is a short questionnaire requiring participants to rate their perceived workload across six likert-type scales which ask questions such as, "How much mental and perceptual activity was required?" (Hart, 2006). The task is easily distributed with participants responding to a series of scales related to constructs enveloped by "workload", followed by comparisons between these constructs – i.e., "which of these two factors represents the more important contributor to workload for the task?" (Hart & Staveland, 1988). These sub-scales include mental, physical and temporal demands, as well as frustration, effort and performance pressures, which can be seen in Figure 2.1 . Hart and Staveland (1988) proposed the TLX after extensive investigation into the six constructs underpinning

workload in most tasks. The comparison part of the experiment asks participants to identify which of the dimensions were most impacted by the task. The TLX assesses workload after the task has been completed, or at intervals throughout the task. In Hart's 2006 review, she notes the wide use of the TLX across a variety of regions and organizations, evidencing this measure as widely applicable to a range of tasks and settings.



RATING SCALE DEFINITIONS		
Title	Endpoints	Descriptions
MENTAL DEMAND	*Low/High*	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
PHYSICAL DEMAND	*Low/High*	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
TEMPORAL DEMAND	*Low/High*	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
EFFORT	*Low/High*	How hard did you have to work (mentally and physically) to accomplish your level of performance?
PERFORMANCE	*Good/Poor*	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
FRUSTRATION LEVEL	*Low/High*	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

FIGURE 2.1: NASA TLX overview from Hart (2006, p. 908) showing the separate domains assessed by the measure and their definitions.

The TLX, as a widely used measure of workload, shows clear strengths in its design. The whole task takes around ten minutes to complete and can be easily completed and calculated online. The task is also able to be completed multiple times without any practice effects. The TLX, like all subjective assessments however, does present several limitations. Primarily, the TLX relies on self report. This method of reporting has long been shown to be subject to bias, misunderstanding, interpretation and memory lapses (Paulhus, 1991). For the TLX, especially in applied uses, social desiriability bias may play a major role in responding, hence limiting this measure (Randall & Fernandes, 1991). Further, the TLX is presented *after* a task has been completed, which means that we only obtain one workload measurement for that time period and participants are forced to rely on memory when being questioned on particular parts of the task. This limitation leads to a range of issues, from memory decay problems, to attribution errors (Schacter, 1999).

Evidently, with the vast amount of research based on the TLX, it appears a valid and reliable measure of workload, and can be applied quite broadly. Despite these advantages, it

is evident that issues of self report still persist and make it difficult to compare individuals or performance across the entirety of the task.

2.2 Physiological Measures

Here I define psycho-physiological measures of cognitive workload as encompassing neural measures, optical measures and cardiovascular measures. Evidently there is a broad range of measures within this section which can be used to assess cognitive workload. Psycho-physiological measures are becoming more readily available, with the development of wearable devices, and in many fields, are the preferred measure of constructs. Psycho-physiological measures allow for a deeper insight into the physiological underpinnings of a behaviour. Compared to self report measures, psycho-physiological measures allow insight into unobservable, and bias-free, latent variables. For example, Schmidt, Decke, Rasshofer, and Bullinger (2017) showed that despite participants reporting increased vigilance after three hours of driving, psycho-physiological measures indicated vigilance was markedly decreased. In comparison with behavioural measures, psychophysiological measures offer greater depth, allowing insight into the mechanisms which may contribute to a behaviour across the span of the experiment. For a full overview of practical applications of psychophysiological measures, see the review by Lohani, Payne, and Strayer (2019).

2.2.1 Eye Tracking & Pupil Dilation

Eye tracking, and more specifically pupil dilation, methods have been proposed as valid and reliable forms of cognitive workload measurement (Biondi, Balasingam, & Ayare, 2020; Kahneman, 1973; Lohani et al., 2019). These measures have been predominantly used in computer science and engineering to provide insights to usability of interfaces (Thorpe et al., 2019). Eye tracking data is often viewed as a measure of the participants state of awareness, which can subsequently infer cognitive workload (Biondi et al., 2020). Fixation time, frequency of fixation change and pupil diameter are all common measures used in eye tracking, which are a behavioural correlate of underlying attentional resource allocation and ongoing cognitive processes (Lohani et al., 2019). In studies by Klingner (2010),

Causse, Peysakhovich, and Fabre (2016) and Krejtz, Duchowski, Niedzielska, Biele, and Krejtz (2018), it was found that participants experiencing high workload had significantly lower gaze fixation times, as well as lower pupil dilation. These results provide evidence of psycho-physiological reactions to increased cognitive workload, and further, evidence such measures as indicators of workload.

Overall, eye tracking and pupil dilation present valid measurement methods of attentional resource allocation, with the ability to model the way in which participants are shifting, and focusing, their gaze. However, apart from gaze time, there is no clear indicator of cognitive workload. Frequent attention shifting is used as the main measure of cognitive workload, however, many other contributing factors can affect results under lower conditions of workload as well. Further, in pupil dilation studies, it is often difficult to measure the size of the pupil consistently and even in cases where it can be estimated, pupil dilation can be affected by other external factors (Lohani et al., 2019). Most importantly however, is that eye tracking measures only provide a correlate of workload rather than linking to the real outcome. This limitation will be discussed further below.

2.2.2 Cardiac Measures

Further psycho-physiological indicators of cognitive workload include cardiac measures such as heart rate variability and blood pressure. Heart rate (and heart rate variability) is also a measure of the sympathetic nervous system, where increased heart rate is associated with increased arousal. Another method of analysis is to asses blood pressure, where increases in systolic pressure are generally associated with increased states of stress (Hughes, Hancock, Marlow, Stowers, & Salas, 2019).

Physiological measures have been used extensively in measuring cognitive workload, with increased heart rate and corresponding with increased workload (Hughes et al., 2019; Reimer & Mehler, 2011). Heart rate measurement allows for precise temporal resolution, with physiological events closely correlating with task events, and has been used in a variety of theoretical and applied settings. Reimer and Mehler (2011) and Mehler, Reimer, and Wang (2011) showed the effectiveness of heart rate measurement as a cognitive workload indicator in driving studies. Further, Hughes et al. (2019) highlight the sensitivity of cardiac measures

in response to varying workload conditions. A further measures discussed by (Hughes et al., 2019) was blood pressure, which was sensitive to increases in cognitive workload, where blood pressure increased with cognitive workload.

Engström et al. (2005) however, showed that despite differences observed in workload in other measures of workload, heart rate appeared to be less sensitive, showing no such difference across conditions. This result speaks to the limits of psycho-physiological measures, as they not only require expensive equipment which can be difficult to distribute, use and analyse, but also require very stable environments where effects are generally small. Another important limitation of physiological measures, which links to a point made earlier about eye tracking measures, is that they provide only a correlate of workload and do not provide an indication of the true outcome. For example, two individuals could perform equally well on a given task, and have even heart rates, yet experience entirely different workload. The inverse may also be the case, where two individuals may differ in heart rate, yet experience the same amount of workload. This is a primary limitation to physiological correlates of workload paradigms.

2.2.3 Galvanic Skin Response

Another psycho-physiological measure which can inform cognitive workload estimates is galvanic skin response (GSR). GSR measures the electrical conductance of the skin through one or two sensors generally attached to the participants hand or foot. GSR measures change in the sympathetic nervous system, as increased conductivity is associated with increased sweating.

Similar to cardiac measures, and closely linked, GSR allows for a physical indicator of cognitive workload as a measure of the sympathetic nervous system. However, GSR. has less temporal sensitivity than cardiac measures, with sweat responses activated slightly later than cardiac responses, however, the two responses are closely linked. There have been several accounts of GSR as a measure of cognitive workload, as seen in work by Nourbakhsh, Wang, and Chen (2013), Nourbakhsh, Wang, Chen, and Calvo (2012) and Shi, Ruiz, Taib, Choi, and Chen (2007) who implemented such measures in laboratory based settings. Further, Engström et al. (2005) used GSR (and heart rate measures) to study workload in a distracted

driving task, as discussed above. Similar to heart rate measures, GSR offers only a correlate of workload, which may not be sensitive to specific task demands, and further, fails to give insight into primary tasks which lack clearly quantifiable outcome measures.

GSR and heart rate may be excellent measures of physiological arousal, however, in workload research, where environments may differ between studies and contexts, they appear practically infeasible. Wearable devices, such as smart watches and heart rate monitors, have made psycho-physiological measures more accessible in recent times, however, there is difficulty in accessing and analysing data (Hicks et al., 2019), as well as privacy concerns over continuous data collection and lack of user control (Motti & Caine, 2015; Thierer, 2015). As these devices, and data, become more accessible, future research should look to combine GSR and heart rate data with other workload measures to form a broader understanding of the construct.

2.3 Neural Measures

Neural measures of cognitive workload have been explored with both electroencephalography (EEG) and functional Magnetic Resonance Imaging (fMRI) techniques showing validity. EEG measures electrical activity at a scalp level. Neural events in higher cortical areas generate electrical impulses which reach the scalp. EEG methods are able to infer active areas of the brain with minimal invasiveness and at a high temporal resolution. fMRI images are formed using magnetic resonance imaging, where non-invasive magnetic radio waves detect changes in blood flow. Oxygen rich blood shows a different magnetic resonance response to blood that is oxygen deficient, and this contrast yields the blood oxygen level dependent (BOLD) signal. The BOLD contrast looks at changes in hemodynamic response (blood flow) to different areas of the brain. This increased blood flow is a correlate of neural activity, which in turn is an indicator of mental activity (Forstmann & Wagenmakers, 2015). Both methods of neurological assessment have associated advantages and disadvantages, but both have been widely used in investigating a variety of phenomenon, including cognitive workload.

2.3.1 Electroencephalography

In cognitive workload contexts, EEG measures are seen as a correlate of activity which shows changes in workload from extrastriate cortex activity. In Gevins et al. (1998) early EEG research, they show an associated increase to frontal theta and decrease to frontal alpha as workload increased. Gevins et al. (1998) show a 95% accuracy for discrimination between high and low levels of workload when observing event related potentials and stress that these changes are likely due to task related difficulty increases. They also proposed that another contributing factor to these changes could be the increase in proportion of cortical resources dedicated to the higher workload conditions. In research by Mühl, Jeunet, and Lotte (2014), which used EEG as a measure of workload, they further this argument, showing the validity of EEG as a workload measure by comparing it to pre-established measures in the domain of attentional workload. Additionally, Frey, Daniel, Castet, Hachet, and Lotte (2016) provide a framework for evaluating user experience using EEG as a measure of cognitive workload with the goal of reducing workload of interfaces.

Despite the strengths of EEG as a measure of workload, it has several limitations. Primarily, the equipment has high set up costs, is expensive and highly sensitive, meaning that experiments need to be conducted in highly controlled environments. Secondly, EEG equipment is less invasive than fMRI, however, is still more invasive than many other measures of cognitive workload, which may interfere with experimental interfaces, particularly in applied settings such as workplaces or driving/aviation simulators. Finally, similar to physiological measures outlined above, EEG provides merely a correlate of workload, without giving an indication of the actual outcome, i.e. research shows that frontal theta waves related to level of workload, however these results do not show resulting workload outcomes.

2.3.2 Functional Magnetic Resonance Imaging

fMRI is another frequently used measure of mental activity which can be used to measure workload. fMRI typically shows high spatial accuracy (within 1mm), yet has lower temporal accuracy (poorer than 1 second). fMRI can be used as a measure of cognitive workload and capacity by assessing neural activity changes associated with changes in load under different conditions in a block of activity. This method not only allows for reasonably

accurate temporal resolution (as the tasks are blocked), but can also provide evidence to areas responsible for differences in workload and capacity. As of yet, there are limited studies using similar methods, which may be due to high experimental costs or the invasive nature of measurement which makes many common cognitive workload tasks, such as driving, inaccessible. These testing environments are inaccessible as fMRI relies on participants lying down in a large scanner and remaining relatively still for an elongated length of time.

fMRI as a form of workload measurement can be useful in understanding which areas of the brain are active during periods of increased workload. fMRI equipment is less accessible, as it is expensive and data requires complicated cleaning and analysis. Whilst fMRI technology shows clear benefits to understand areas of the brain responsible for attentional control, attentional shifts and identifying the neural regions involved in completing complex tasks, the setup and analysis costs, as well as the nature of the measure as a correlate of behaviour, means that it still has drawbacks as a widely used cognitive workload measure.

2.3.3 Functional Near Infrared Spectroscopy

Functional near infrared spectroscopy (fNIRS) is another neural measure of cognitive workload. fNIRS uses near infrared spectroscopy (an optimal imaging technique) to measure changes in hemoglobin levels in the brain. This technique is similar to fMRI, however, has lower spatial resolution, relying on measurements from the scalp. fNIRS is often combined with other neural measures (such as EEG) to evaluate neural and hemodynamic responses to given stimuli (Aghajani, Garbey, & Omurtag, 2017). In workload measurement, similar to other neural workload measures, changes in the frontal cortex (such as increased hemodynamic levels) are related to changes in cognitive workload (Causse, Chua, Peysakhovich, Del Campo, & Matton, 2017; Herff et al., 2014). Although fNIRS techniques have proven useful in evaluating workload, the measure is limited in similar ways as EEG and fMRI. fNIRS is promising for future research in cognitive workload measurement, as it is temporally accurate, less invasive and more accessible than other neural measures – however, provides less depth to assessment. Furthermore, similar to the limitations outlined above, fNIRS provides only a correlate of performance, rather than assessing cognitive workload outcomes.

2.4 Dual Task Measures

Dual task measures of cognitive workload are based on the theory of a limited capacity system. Dual task paradigms specify that an individual completes a main task (or possibly several main tasks) and a secondary task. The main task/s requires a portion of the individuals cognitive capacity to be occupied. The residual capacity is then dedicated to the secondary task. Therefore, if workload demands from the main task are high, performance on the secondary task (the cognitive workload measure) would suffer. Similarly, if main task demands are low, a greater amount of attention can be allocated to the secondary task.

The secondary task in a dual task measure is generally simple, requiring minimal attention and minimal response effort, to ensure that the secondary task does not impact main task performance. The primary example of cognitive workload measurement in a dual task framework is the Detection Response Task (ISO:17488, 2016). The DRT requires participants to detect a salient signal and respond by pressing a button as fast as possible. If the main task has high workload demands, response times will be inhibited (Strayer et al., 2013), consequently providing a simple measure of residual capacity. Dual task measures have shown effectiveness at quantifying cognitive workload in a variety of settings (Engström et al., 2005; Stojmenova & Sodnik, 2018; Strayer et al., 2013). With a simple procedure, minimal task demands/distraction, minimal equipment costs, quantitative data and high sensitivity to workload changes (Stojmenova, Jakus, & Sodnik, 2017), dual task measures such as the DRT are able to be used in a variety of contexts to answer a range of cognitive workload questions. Similar to other cognitive workload measures; dual tasks do have several limitations. These limitations include the ability to trade off performance between tasks (however this can be resolved through analysis of the main task) and a discontinuous measure of workload as signals occur at discrete intervals. This limitation is minimized as responses are required frequently (every three to five seconds in the DRT), and, by using a blocked design (similar to fMRI studies), can be further reduced.

Whilst researchers could choose to focus purely on main task performance, there are shortcomings which can be addressed through dual-task methods. Primary task performance typically decreases with increased demand, or workload. The additional "workload" task is useful however, when the main task cannot provide reliable performance measures. For

example, performance may fluctuate throughout a driving scenario, however, it is difficult to quantify driving performance. Similarly, many tasks may only provide a single outcome measure, which can be difficult to interpret over a long block of activity where workload may fluctuate throughout the block. Including a secondary task which is simple and provides a constant and reliable measure of workload allows inferences in such tasks to be possible. Further, including a dual task allows us to compare workload across contexts and tasks where the performance metrics may be diverse.

Overall, dual-task measures present one of the most accessible forms of direct cognitive workload measurement. Dual-task measures may not account for the architecture producing behaviour (as neural measures do), nor do they account for sympathetic effects (as physiological measures do), however, it could be argued that in order to *evaluate* cognitive workload, these factors are not vital to understanding the underlying workload induced in the context of interest. Finally, unlike neural and physiological measures discussed, dual-task measures provide more than a correlate of workload, instead, results highlight the behavioural outcomes of varying levels of cognitive workload.

2.5 Detection Response Task

Detection response tasks were first presented as workload measures in 1999 by Van Winsum, Herland, and Martens (1999) in their paper "The effects of speech versus tactile driver support messages on workload, driver behaviours and user acceptance". The authors termed the task the "peripheral detection task" (PDT), where participants were required to detect and respond to a signal in the periphery. The DRT terminology was later introduced by Engström et al. (2005) and shares many close similarities with the PDT. The DRT was standardized under ISO:17488 (2016) in an attempt to unify the various dual-task cognitive workload measures which are methodologically similar to the DRT (such as the PDT as in Van Winsum et al. (1999) and tactile detection task in Diels (2011)).

The DRT operates as an additional task in dual task paradigms. Resources required to attend to the additional task – the DRT – are limited to the residual attentional resources available. As a result, the DRT is an indicator of residual capacity. The DRT imposes a

minor demand on resources (Biondi et al., 2020) due to its simple nature, however, results capture the availability of residual resources.

If an individual is doing a highly complex task, they will have limited residual resources. For example, in a driving scenario (a highly complex task), individuals residual resources are limited, consequently making it difficult to complete tasks such as remembering words or completing simple maths problems (Strayer, Turrill, et al., 2015). In a simultaneous DRT trial of the present paradigm, the individual will have less resources to detect and then respond to the signal. Consequently, DRT responses will be slower on average. Alternatively, an individual who is completing a simple task, with minimal effort, would have a large amount of resources leftover. Here, the individual would have a greater amount of resources available to detect and respond to a DRT signal, meaning that responses would be faster. These trends are consistently shown in workload research, and provide a key result throughout this thesis – as shown in Figure 2.2; slow responses correspond with high workload, fast responses correspond with low workload.

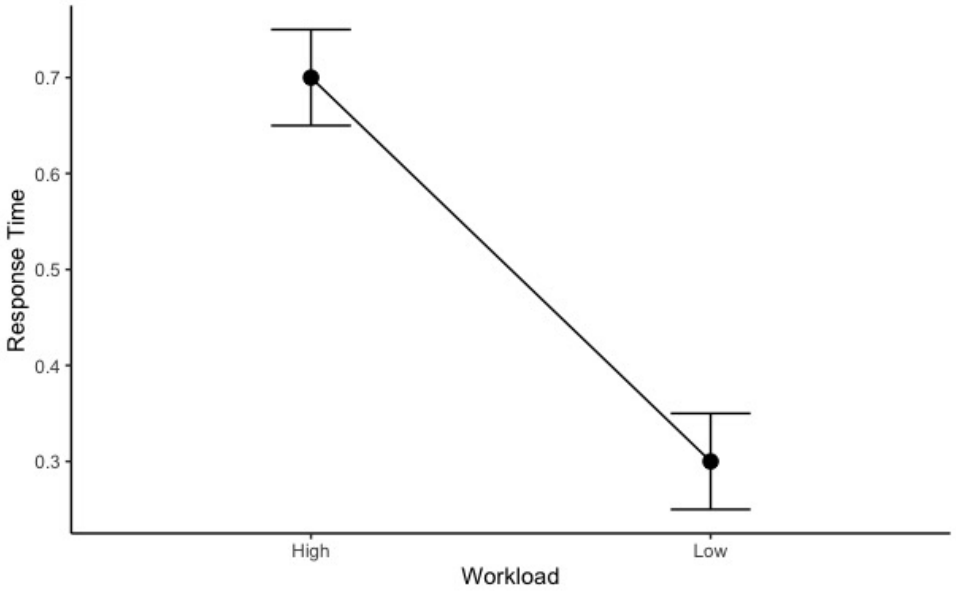


FIGURE 2.2: Theoretical results for a DRT study. In conditions of high workload, participants average DRT response times are slower. In conditions of low workload, response times are faster.

2.5.1 Driving and the DRT

Van Winsum et al. (1999) showed the effects of increased workload in distracted

driving scenarios as a first pass at using quantitative workload measures in driving environments. Their task focused on assessing workload increases induced by difficult driving scenarios and assistance messages. The driver assistance messages were given as warnings via speech or tactile modes. The researchers found that workload increased with more difficult driving scenarios (for example response times were low on a straight road and higher at stop signs or when forced to brake suddenly). Furthermore Van Winsum et al. (1999) showed that misses to the PDT also increased in these more difficult scenarios – again indicating an increased workload. The researchers showed the sensitivity of the PDT as a selective attention measure with response times increasing during, and after, the warning systems had been activated. This finding is a prevalent theme in cognitive workload measurement literature, as although the system enabled safer driving, it had a trade-off with cognitive workload. That is, to process and utilise the system, required more cognitive resources, but ultimately lead to safer driving.

Similar driver distraction research has been undertaken by David Strayer and colleagues from the Applied Cognition Laboratory at the University of Utah. In their initial papers, Strayer and Johnston (2001) used a braking task, which measured the response times of drivers to simulated traffic signals. The researchers showed that in multitasking scenarios (such as using a mobile phone), participants had significantly slower reaction times and were much more likely to miss the simulated traffic signals. Similarly, in Strayer et al. (2003), the researchers showed the differences in detection ability between cell-phone conversations in comparison with radio broadcasts or audio-books. Results indicated that participants again had a greater probability of missing simulated traffic signals and slower response times. Further studies investigated differences between older drivers and younger drivers in relation to their dual-tasking ability (Drews, Pasupathi, & Strayer, 2008; Strayer & Drew, 2004), and differences between drunk drivers and distracted drivers (Strayer, Drews, & Crouch, 2006). These brake response time studies have been highly influential in policy change regarding cell-phone use, and provided a deep insight into driver distraction, however, they only accounted for "workload" at critical periods (i.e. stopping at traffic signals). It is clear however, that results could have been impacted by a task switching cost, rather than an overall workload increase. A more effective measurement of the workload induced by cell-phone interaction should also account for time when the driver was not engaged in a critical driving behaviour, such as braking. The DRT is able to do this.

Strayer et al. (2013) introduced the DRT to their driver distraction research. Since then, the DRT has been used successfully across a variety of distracted driving studies. Using a baseline technique, the Applied Cognition Lab has developed a method of comparing distractions in the automobile setting. This technique has been used to account for the effects of a variety of distractions on driving including talking on the phone (Strayer et al., 2013), interacting with in vehicle information systems (Strayer et al., 2019) and interacting with smart assistants (Strayer et al., 2017). This research has been crucial for driver safety and policy development, and highlights the usefulness and applicability of the DRT.

Further, on the usability of the DRT, there are several examples of research using varied DRT signal modalities, including visual, tactile and auditory, to be usable in alternate environments. For example, in some environments, a tactile stimuli would be less distracting and intrusive than a visual signal. There have been attempts to differentiate these modalities (see Conti et al. (2012); Merat and Jamson (2008); Stojmenova et al. (2017) for full reviews), but all appear to be sensitive to cognitive workload changes. Recently, Biondi et al. (2020) showed that using the DRT in psychological experiments does add minor workload to the user (the researchers measured this impact using pupil dilation measures in a between subjects n-back task). Additional workload added from the DRT is both intuitive and necessary, as adding another task should increase the workload of a participant, however, adding an extra task enables us to understand the magnitude of alternate workload inducing factors – such as the additional workload from texting whilst driving, or the impact of increased task difficulty on workload.

Aside from driving literature, there is little evidence of DRT use in alternate settings (Innes, Evans, et al., 2020). Such a simple and applicable task could be easily applied to a variety of other settings to enable a deeper view into the underpinnings of task performance, distraction and cognitive workload – my thesis aims to address this in part. The following section details some of the studies which validate the use of the DRT outside of driving contexts, for example in lab settings, such as task switching exercises, and alternate applied settings, such as helicopters and forklifts, findings of which are important for assumptions I make throughout this thesis. The majority of these studies aim to answer applied questions only.

2.5.2 Using the DRT in Alternate Settings

ISO:17488 (2016) for DRT outlines methods of the DRT in evaluating driver cognitive workload. These methods are closely followed throughout this thesis when referring to DRT methodology. The only significant deviation is that all thesis methodology is unrelated to driving paradigms. The DRT methodology requires (at minimum) a dual-task paradigm, where the DRT is an additional task. Consequently, participants attention is focused on the main task – the task which occupies the majority of cognitive workload. Task preference can be specified to participants (for example "performance in the driving task is most important in the current scenario"), however is often ignored by researchers and is shown to have little effect on results (Conti, Dlugosch, & Bengler, 2014). ISO:17488 (2016) specifies that the experimental task is the driving task, however, a limited number of other experiments have used DRT dual-task paradigms in alternate settings (Biondi et al., 2020; Engström et al., 2013; Gross et al., 2018; Innes, Howard, et al., 2020; Thorpe et al., 2019; Xie et al., 2016; Young et al., 2013). To test differences between conditions, these tasks must remain engaging and difficult in order to observe workload changes.

The studies outlined above all highlight the usefulness of the DRT methodology outside of driving literature. In a closely related context, Gross et al. (2018) used the DRT to measure the impact of alternate monitors on cognitive workload in forklifts, with results indicating the sensitivity of the DRT. Similarly, Thorpe et al. (2019) used the DRT as a cognitive workload measurement tool in relation to gaming user interfaces in order to assess differences in usability. Thorpe et al.'s (2019) methodology used the DRT in a tightly controlled environment, with a simple cognitive task - a tracking task. Similar experiments which aimed to use the DRT in controlled laboratory settings include studies by Engström et al. (2013) and Young et al. (2013). These examples used an n-back task rather than a driving paradigm to manipulate workload demands. They showed that the DRT was sensitive to the difficulty of the task, with greater n-back corresponding with increased DRT response times. Similarly, Biondi et al. (2020) used an n-back task and pupil diameter measures, to evaluate the effect of the presence of the DRT on workload. Finally, Xie et al. (2016) highlighted the usefulness of the DRT in assessing group workload, by evaluating DRT results for individuals within a group setting. Xie et al.'s (2016) group experiment took place inside a military special vehicles context, however, not all participants were completing

the driving task, but rather a variety of tasks related to the specified mission. These results further highlight the usefulness of the DRT at not only quantifying individual workload, but extending the paradigm to a group context to evaluate interpersonal factors on cognitive workload.

Chapter 3

The DRT-MOT Design: A validation study

One major goal of this thesis is to extend the applications of the DRT. In Chapter 2, I show a novel way of applying the DRT to measure cognitive workload induced by the Multiple Object Tracking Task (MOT). Further, I show how this task can also be distributed online. Previously, DRT experiments have been for the most part conducted in driving environments, such as simulators or controlled driving course experiments. Prior to extending the applications of the DRT within the DRT-MOT task framework, it is important to evaluate the task under different experimental manipulations to provide evidence for the reliability, and validity, of the DRT in contexts outside of typical driving scenarios. This is important, both in terms of this thesis and in the scope of broader literature, to portray the usability and reliability of the DRT, where in some contexts, the interrogated task may be less cognitively demanding and so workload ceiling may not be reached. This could impact the sensitivity of the measure.

In this chapter, I explain the details of an experiment which evaluated DRT signal modality. Two other experiments were also conducted which evaluated the DRT in reference to another measure of cognitive workload (the NASA TLX) and against a measure of an alternate construct – vigilance (the PVT) [1]. These subsequent experiments are limited to simple methodology and overall trends to restrict the length of this chapter. From the experiments in this chapter, I aim to show both the reliability of the DRT as a cognitive workload measure and the sensitivity of the measure to workload change.

3.1 Experiment 1: DRT Signal Modality Comparison

Research has shown our cognitive capacity limitations are shared across various modality channels (Diederich & Colonius, 2004). We are able to focus our attention on one modality source, however, this does not increase our overall capacity. For example, if we focus only on our auditory channel, we are still limited to listening to only one conversation. Many cognitive tasks, such as n-back and operation span tasks are presented through various modalities (usually visual or auditory). Despite different stimulus presentations and

[1] Results (in the form of JASP outputs) from these experiments can be found online at `https://osf.io/ayp6d/`

associated processing speeds relevant to each modality, trends hold across these conditions (Stojmenova et al., 2017).

Distracted driving studies which use the DRT as a form of cognitive workload measurement have employed several signal modality presentations (Cooper, Castro, & Strayer, 2016; Stojmenova et al., 2017; Stojmenova & Sodnik, 2018). This signal could be tactile – a vibration elicited by a vibration motor attached to the skin; auditory – a salient sound elicited in the display; or visual – a salient light source delivered in the periphery or within the environment. As outlined in Chapter 2, increasing workload leads to increased response times. However, altering signal modality can also effect response times. Stojmenova et al. (2017) and Engström et al. (2005) showed differences in response times across DRT signals (in Engström et al. (2005), the TDT and PDT - see Chapter 2 - were evaluated). Both studies found that all signal modality presentations were sensitive to workload changes. Trends showed that the tactile signal was responded to fastest in some conditions, and visual signals appeared the most sensitive to change between conditions, however, results were not conclusive.

Of the studies which use the DRT, the majority of these are in practical and applied scenarios, such as driving studies (Stojmenova et al., 2017; Stojmenova & Sodnik, 2018). Importantly, in any DRT experiment, the modality of the signal should be sensible given the task. For example, in Strayer et al. (2013), drivers were asked to engage in conversations with a passenger or via a hands free mobile phone. In this design, using an auditory DRT signal may confound the affects of the incoming conversation. Instead, researchers used a visual DRT signal to limit any confounding effects.

In Innes, Evans, et al. (2020), the authors show differences between two versions of the DRT-MOT paradigm. In the initial experiment, the DRT signal was delivered through the tactile modality, whereas the secondary study used a visual signal. Results showed that the DRT was sensitive to workload changes induced by varying MOT difficulty, however, it was also noted that the visual signal may have been distracting for participants. In Innes, Evans, et al. (2020), a difference in response times was observed between the visual and the tactile conditions, with visual having much slower response times. The paper is limited as these differences could not be analyzed due to different samples of participants, whose tasks

differed slightly. These differences may be the result of a preference for tactile signal, or may be due to inhibited processing when the DRT and MOT signals were both visual.

The current experimental procedure was based on Innes, Evans, et al. (2020) DRT-MOT methodology, where the MOT is used as a controlled manipulator of cognitive workload and the DRT is used as the cognitive workload measure. DRT response times generally increase as cognitive workload increases, essentially indexing workload (Cooper et al., 2016; Stojmenova & Sodnik, 2018; Strayer et al., 2013). In easy conditions, DRT response times are lower, and in hard conditions, response times increase. This phenomena can be thought of as the amount of attention and concentration given to a task dependent on the task difficulty, where your cognitive resources "run out" as the task becomes harder or requires more concentration. The DRT captures this depletion of resources in the addition of a simple stimulus detection task.

In typical studies of workload using DRT measures, the simultaneous task is generally highly complex (such as driving) yet often limited in control (due to this complexity). The MOT used in the current design differs from these prior studies in that it is less cognitively, and physically, demanding than driving – requiring only visual tracking of stimuli and button press responses. Furthermore, the MOT has shorter periods of activity and is able to be tightly controlled – as the number of dots to track indexes the task difficulty and all other experimental factors are kept constant (such as the speed of the motion, the size of the objects and the length of tracking time). The current experiment aimed to show the sensitivity of the DRT to workload change in these tightly controlled tasks synonymous with cognitive psychology research. Specifically, the MOT presents a highly repetitive task where no responding is required during the tracking (workload) period, which is divergent from the typical DRT-driving literature. Further, it is important to evaluate varying stimulus modality presentations across this task, to establish reliability in online distribution, where the use of a physical DRT is not possible. Evidently, two main effects *should* be observed, similar to results from Innes, Evans, et al. (2020): MOT performance decreases with difficulty – showing that participants find the task more difficult with more objects to track; and DRT performance decreases with difficulty – showing that as participants work harder in the MOT, their attention is detracted from the DRT, causing performance to drop. This co-occurring result pattern is common in DRT designs and throughout this thesis.

Following Innes, Evans, et al. (2020), I aimed to investigate the differences in both DRT and MOT results given alternate DRT modality signal presentations within DRT-MOT paradigm. Furthermore, I aimed to investigate the differences between the "physical" DRT signal (the light or vibration) and the "virtual" signal (a visual signal given on screen alongside the cognitive task), to test the reliability of the online version of the task. Given the results of previous literature, it was hypothesized that all DRT signal modality presentations would be sensitive to changes in cognitive workload, with response times increasing as MOT difficulty increased. Secondly, it was hypothesized that there would be a difference in DRT response times between DRT signal modality, with tactile predicted fastest and virtual slowest in line with previous research from Engström et al. (2013) and Stojmenova et al. (2017). Finally, it was hypothesized that DRT signal modality would have no affect on MOT results.

3.2 Method

3.2.1 Participants

Participants were 22 undergraduate psychology students of the University of Newcastle who were reimbursed with course credit. All participants had normal or corrected-to-normal vision and were able to read English. Participants completed the study in lab. A total of 4 participants were removed from the analysis due to missing data (i.e. not completing the experiment) or technical faults (two online files did not save data).

3.2.2 Tasks

The method was based on Experiment 2 of Innes, Evans, et al. (2020), with differences only in the DRT signal modality presentation and overarching design. Participants were required to complete two simultaneous tasks; the MOT task and the DRT. The design was a 3x2 within subjects design. No instructions were given to specify the preference of DRT or MOT tasks to participants (see Conti et al., 2012, for a full review of DRT instructions). The MOT had two levels of difficulty, indexed by the amount of dots to track (2

or 4). The general MOT administration was identical to that used in Innes, Evans, et al. (2020).

The DRT generally adhered to ISO standardization (ISO:17488, 2016). In the DRT, there were three signal types' visual, tactile and virtual. Participants were required to respond to a short signal (1 second), which could be tactile, visual or virtual. The tactile signal was a vibration elicited by a motor attached to the participants shoulder - the same as in Experiment 1 of Innes, Evans, et al. (2020). The visual signal was a red LED light which was attached to the participants head via a velcro strap so that it was seen in the periphery, as used in (Strayer et al., 2013). The virtual signal was included in the MOT display. For the virtual condition, a red frame appeared around the display, similarly following ISO:17488 (2016) standards. This signal was different from Innes, Evans, et al. (2020), as the signal was much larger and encompassed the MOT motion area (as seen in Figure 3.1), enabling participants to maintain focus on the tracking area for the MOT.

FIGURE 3.1: Screenshot of the virtual DRT signal condition. The signal encompasses the multiple object tracking area, so that an object's motion is reflected before reaching any area where the frame could appear.

Participants responded to the signal in the tactile and visual conditions by pressing a button attached to the index finger of their non-dominant hand. In the virtual condition, participants pressed the space bar to respond to the stimulus onset. The DRT signal lasted for 1 second, unless the participant responded before the onset time had elapsed. DRT signal onset occurred in cycles, with the interval between cycles randomly distributed between 3 and 5 seconds. Participants were instructed to respond as quickly as possible to the signal before

the next signal occurred. Responses made before the next occurrence of a signal were deemed "hits" and a failure to respond before the next signal, or within 2.5 seconds, was deemed a "miss". Second (and subsequent) responses entered before the onset of the next signal were deemed "false alarms", however, as per (ISO:17488, 2016), these were not included in the analysis. Response times were measured as the time taken to respond following the onset of the current signal. Participants completed four blocks (two of both 2 dots and 4 dots to track) of MOT for each condition of DRT signal, for a total of 12 blocks.

3.2.3 Procedure

Participants were first briefed on the experimental setup and the different equipment used for each DRT signal type. Participants were then instructed on the MOT task and shown examples of how the task was presented. Participants were shown how to respond to the DRT. They were also given several practice trials of each condition of DRT.

Each DRT signal condition was blocked by design so that participants completed four blocks of MOT for each signal type without changing equipment. The order of DRT signal presentation was randomised between subjects. The order of MOT difficulty was randomised between subjects and between DRT signal conditions. Participants were given two practice trials of both levels of MOT difficulty (2 dot and 4 dots) at the start of each block of DRT signal type. Each MOT block consisted of seven trials. Within each block, all the trials used the same number of dots to track: either two or four. Each of these levels of difficulty was used for two blocks, giving a total of 14 MOT trials for each cell of the design, a total of 84 trials. Note that these 84 trials refers to the MOT, as the number of DRT trials would change due to the randomisation of inter-trial intervals. Also note that each MOT trial refers to each period of tracking the objects. Participants would make five decisions (target or not a target) for each MOT trial. On average, blocks took five minutes to complete. Participants were able to take short breaks between blocks. Changing the equipment between DRT signal conditions took $\sim$ five minutes. The total time taken to complete the experiment, given setup, testing and break time, was between 1-1.5 hours.

3.2.4 Thesis Analysis Overview

Analysis was conducted using the "BayesFactor" and "bayestestR" packages in R, with default priors. I used the "bayestestR" package to evaluate Bayesian inclusion probability (written as $BF_{inclusion}$) for main and interaction effects of each predictor in Bayesian ANOVAs. The inclusion Bayes factor quantifies the change in inclusion probability from the prior to the posterior, which can be interpreted as the amount of evidence from the data for including a predictor (van den Bergh et al., 2020). The prior inclusion probability is the sum of prior model probabilities of all models containing the predictor. The posterior inclusion probability is the probability that a predictor is included in the model, computed as the sum of all model probabilities containing a given predictor after seeing the data (van den Bergh et al., 2020). For Bayesian ANOVA results, I will refer to $BF_{inclusion}$ as the probability of data under a model containing a given predictor, compared to models without this predictor. Greater $BF_{inclusion}$ indicates greater evidence for the effect of a predictor on the dependent variable of interest. $BF_{inclusion}$ below 0.3 indicates evidence for a null effect of a predictor – values below one will be presented as fractions in the ANOVA tables. Bayes factors above 3 indicate evidence for the inclusion of a predictor. Any Bayes Factor over 1000 will be referred to as "$BF_{inclusion} > 1000$". Values that indicate strong evidence for or against predictors (i.e. > 3 or $< 1/3$ respectively) are shown in bold in the analysis tables.

All results in following chapters will first be discussed relating to trends and figures, followed by statistical support of trends from the Bayesian analysis (as discussed above), and then post-hoc analyses or further analyses conducted. DRT-MOT results are always shown across two main graphs – one showing MOT results and the other showing DRT results (as well as any accompanying sub-figures). $BF_{inclusion}$ factors are shown in tables, where each column indicates a different dependent variable (such as DRT response time), and each row shows the effect of a given predictor (such as difficulty). For post-hoc analyses, Bayesian t-tests comparing conditions will be referred to as either evidence in favour of a difference (i.e. the alternative hypothesis – BF_{10}) or evidence in favour of a null difference (BF_{01}) between conditions. I refer to Jeffreys (1961) for interpretation of Bayes factors in favour of the null and alternative hypotheses.

For all experiments, DRT responses under 0.1s were excluded. DRT responses over

2.5s were classed as "misses". Secondary and subsequent DRT responses (prior to the onset of the next DRT trial) were classified as "false alarms". This is standard DRT procedure (ISO:17488, 2016). Participants with under 50% accuracy in *any* MOT condition were excluded. Participants with miss proportions over 50% in the DRT were excluded. Participants with more than 100 false alarms in the DRT were excluded. This false alarm exclusion was most commonly due to a technical error in some experiments where the DRT did not record properly – likely a browser related issue. Poor performance exclusion was defined as DRT miss proportion greater than 50% or MOT accuracy less than 50% for the lowest difficulty condition in each experiment.

3.3 Results

The study was treated as a 3x2 within subjects design, with three levels of DRT signal modality (tactile, visual and virtual) and two levels of MOT difficulty indexed by the number of dots to track (2 or 4). For the MOT, analysis included dependent variables of MOT RT and proportion correct – for correctly identifying targets and rejecting non-targets. DRT results analysed were mean DRT RTs and proportion of misses. Two-way Bayesian ANOVAs were undertaken for each of these measures, with results presented in Table 3.1.

$BF_{inclusion}$	DRT RT	DRT Miss	MOT RT	MOT acc
difficulty	17/25	19/50	**> 1000**	**> 1000**
modality	**2/25**	**44.82**	**160.84**	**3/20**
difficulty:modality	**1/25**	**13/50**	47/50	**1/10**

TABLE 3.1: $BF_{inclusion}$ factors across dependent variables (columns) for each predictor (rows). $BF_{inclusion}$ with sound, or greater, evidence are shown in bold. $BF_{inclusion}$ shown as fractions represent evidence for null effects of the given predictor. $BF_{inclusion}$ greater than three represent evidence for the effects of a given predictor, whilst $BF_{inclusion}$ less than a third represent evidence against the effects of a given predictor.

The MOT was performed well, with an average MOT proportion correct of .80 (SD = .13) and MOT RT of .99s (SD = .27). The change in performance across levels of MOT difficulty can be seen in Figure 3.2, which also highlights the minimal impact of modality on MOT performance. Mean proportion correct declined as difficulty increased, however, there seemed to be no effect of modality on accuracy. Similarly, mean MOT RT increased with difficulty, but again there appeared to be minimal effects of modality. Bayesian ANOVAs

confirmed these trends, as shown in Table 3.1, which indicated strong evidence for the effect of difficulty on MOT accuracy, and strong evidence for the effects of difficulty and modality on MOT RT. Evidence was also shown for null effects of modality and interaction on MOT accuracy, whilst ambiguous evidence was shown for the interaction having an effect on MOT RT. These results indicate that higher difficulty lead to lower accuracy and higher MOT RT. Furthermore, modality appeared to have no affect on participants tracking ability, yet did appear to affect participants MOT RT indicating that participants responding was impaired in this condition (possibly by the alternate button press or the appearance of virtual signal during the MOT interrogation phase).

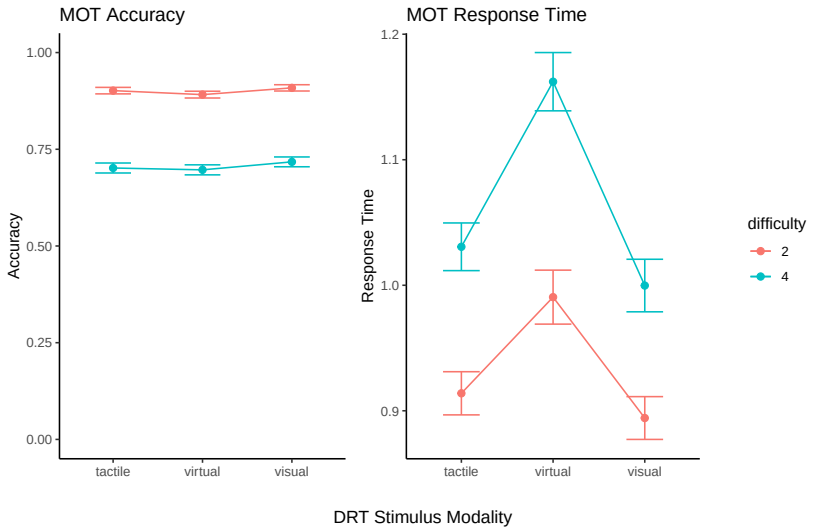


FIGURE 3.2: Performance on the MOT across conditions and difficulty. *Left Panel*: Accuracy in the MOT across levels of difficulty for the three modality conditions. *Right Panel*: Mean response time in the MOT decision phase across levels of difficulty for the three modality conditions. Error bars shown are standard error.

The DRT was also performed well, with an average DRT RT of 0.55s (SD = 0.13) and average DRT miss proportion of .09 (SD = .09). The change in DRT performance across levels of MOT difficulty can be seen in Figure 3.3. Furthermore, Figure 3.3 highlights the differences of modality on DRT performance. Mean DRT RT increased as difficulty increased, however, there seemed to be no effect of modality on RT. Mean DRT miss proportion seemed to increase with difficulty, and showed some effects of modality, with virtual modality showing a slightly higher miss proportion. As shown in Table 3.1, Bayesian ANOVAs confirmed these trends with evidence against the effects of modality, or an interaction, on DRT RTs, and ambiguous evidence for the effects of difficulty. This indicates that DRT signal modality

had no effect on participant DRT RTs, and more evidence is needed to qualify the effects of difficulty. For DRT misses, the Bayesian ANOVA indicated that DRT signal modality did have an effect on misses, whereas evidence was ambiguous for the effect of difficulty (with evidence against an interaction effect). This result indicates that the modality of the DRT may have had an effect on miss proportions, with post-hoc Bayesian t-tests indicating differences between tactile conditions in comparison with virtual and visual conditions (Virtual vs Tactile, $BF_{10} = 13.04$, Tactile vs Visual, $BF_{10} = 3.81$), and ambiguous evidence found for a difference between visual and virtual conditions ($BF_{10} = 0.52$). This indicates that misses were lowest in the tactile condition. This result is in line with streams of information literature, as the tactile modality was likely perceived more efficiently due to the low processing demands from this modality channel (in comparison with the two visual DRT stimuli which are presented in the same modality as the main task).

Additionally, the between subject factor of participants was analysed for DRT RTs. In the Bayesian ANOVAs, the participant factor is included in the null model, however, interpretation of this effect is important to the modality hypothesis, which assumes there will be no difference across modality. $BF_{inclusion}$ indicated strong evidence for the effect of participants ($BF_{inclusion} > 1000$), indicating evidence for a difference between conditions. This result highlights the reliability of the DRT across modalities.

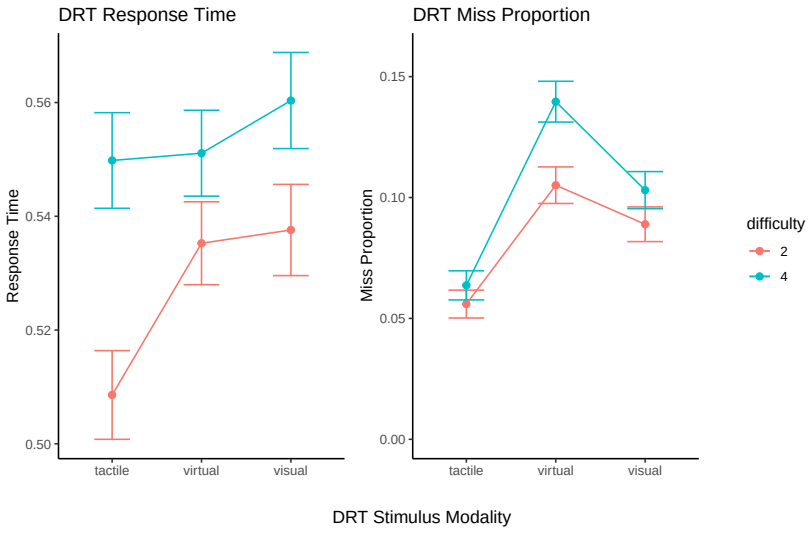


FIGURE 3.3: Performance on the DRT across difficulty and conditions. *Left Panel*: Mean DRT response time across levels of difficulty for the three modality conditions. *Right Panel*: Mean proportion of misses in the DRT across levels of difficulty for the three modality conditions. Error bars shown are standard error.

3.4 Discussion

The current study evaluated whether the effects of DRT signal modality would effect results in the DRT-MOT paradigm. There were three DRT signal presentations (two were the same modality): visual, where a light was shown in the periphery; tactile, where participants received a vibration; and virtual, where a visual signal was presented on the same screen as the MOT. It is important to note that the virtual signal required a different response - a key press - than the visual and tactile signal - a button press. The MOT task, and all other factors of the DRT task, were kept constant between conditions. Comparisons of DRT signal modality allow us to further validate the DRT-MOT design as a sound form of workload measurement and a reliable indicator of performance.

In line with previous studies, a large effect of MOT difficulty was shown on DRT RT and misses, with DRT RT and misses increasing with difficulty. In line with my hypothesis, DRT signal type had almost no effect on MOT performance, with task accuracy and MOT RT stable across conditions. There was however, evidence for differences between signal conditions in DRT RT and DRT misses. This difference was specifically shown in the virtual condition for DRT misses and in the tactile condition for DRT RT. This result was in line with the hypothesis that modality would impact DRT RT due to the different levels of processing required. However, the effect on DRT omissions in the virtual condition was surprising. Despite the differences between DRT stimulus types, there were strong relationships observed between stimulus types, showing that different signal presentations are likely to be measuring the same construct - cognitive workload.

Results from the DRT showed that tactile DRT signals were responded to fastest, which is in line with previous literature (Stojmenova et al., 2017). Furthermore, there was no large difference between visual and virtual signals, however there was a trend in line with expectations where the virtual signal appeared to be responded to slower than the visual signal, and with greater omissions. This could be the result of the virtual signal requiring a different response than the other signal presentations and/or from featuring in the same display as the MOT. As the stimulus type did not appear to affect MOT results, it is likely that the above reasons were the underlying cause of this difference rather than any underlying cognitive mechanisms. The similarity between these stimulus types is an improvement from

earlier experiments (Innes, Evans, et al., 2020), where it was noted that the stimulus was difficult to perceive due to its low salience and small presentation size. Furthermore, results from the MOT indicated that the DRT modality had no effect on MOT responding, showing that the virtual DRT signal did not distract participants from the MOT. These results are promising for the DRT-MOT task, as it shows the reliability of the MOT task under different conditions which speaks to the flexibility of the DRT. The DRT again showed sensitivity to workload change, as observed in Innes, Evans, et al. (2020), showing differences between conditions of MOT difficulty. This experiment provides further validation of the DRT-MOT design, as not only was there no impact of responding on the MOT, but the DRT was just as sensitive compared to other "physical" DRT signals (and responses). Furthermore, when using an online design (i.e. the virtual condition), it is important to continue using the more salient version of the virtual DRT signal (as used in this task) which had no effect on MOT results and showed limited differences from the other "hardware" DRT signal presentations.

The implications of these findings are significant for expanding the use of the DRT-MOT platform to online environments. By providing evidence that stimulus presentations indicate the same trends, we can have more confidence in extending the task to broader environments. These environments include testing interfaces or displays, such as in Thorpe et al. (2019), where we can be confident that the DRT signal is not affecting main task performance. Furthermore, in showing the validity of the virtual signal, it allows for a broader distribution of the task online, using systems such as MTurk to collect data.

Results from the current experiment support the theory of a joint modality capacity, where we are limited overall in our cognitive capacity across modality channels. This means that if we are completing two tasks in different modality channels, for example visual and tactile, we are still limited in a similar way to if we were completing two tasks in the same modality channel (for example visual and visual). These results support other DRT modality studies such as Engström et al. (2013); Merat and Jamson (2008); Stojmenova et al. (2017) and Stojmenova and Sodnik (2018) who have conducted similar studies in applied environments. Evidently, with results supporting similar research, the DRT-MOT paradigm appears to be valid in assessing cognitive workload in lab-based environments.

The differences in MOT RT between stimulus presentation conditions should be evaluated if MOT response timing is important for future analysis, such as modelling MOT

decisions. This difference is likely due to the alternate responses, i.e. key presses for the virtual condition compared to button presses for DRT responses. Further, these differences may have the most effect during the interrogation phase, where dual responses are required. Biondi et al. (2020) showed that the addition of the DRT can inhibit performance on the other task, however, in their study, researchers used an *n*-back task as the workload inducing task. As both the DRT and the *n*-back require discrete responses, issues arise from response switching, which may also be the case in the present study. This is overcome in future DRT-MOT designs, as DRT responses are only required during the MOT tracking phase, thus, there are no overlapping response periods. Consequently, the DRT workload indicator is not affected by MOT responses, and DRT responses do not distract from MOT responding.

The current study is limited in that I only use two levels of MOT difficulty for comparison and do not evaluate auditory signals (however, in Thorpe et al. (2020), we see limited effects of auditory stimulus presentation compared to visual). Future research could evaluate more levels of the MOT, including a baseline condition, to assess the overall change in workload between "no-load" and the workload induced by the MOT. Future research could also evaluate whether results from the DRT correspond with other workload measures. In another study, I tested this aspect, with results discussed below. Furthermore, I also tested whether results from a measure of a different construct differed from DRT results.

3.5 Further Tests of Validity

In addition to Experiment 1, two further studies were conducted in the DRT-MOT framework to examine the validity of the design. These experiments used the DRT-MOT framework and compared **1**) a well established subjective measure of cognitive workload – the NASA-Taskload Index (TLX) and **2**) a well established measure of fatigue - the PVT – with the DRT. Both studies used the DRT-MOT as explained in Innes, Evans, et al. (2020). Results from both studies are included in the form of JASP outputs (JASP Team, 2019) and can be found along with analysis scripts and data at `https://osf.io/ayp6d/`.

In the study comparing the TLX to DRT, participants completed high and low difficulty blocks of MOT. Workload was measured via three methods (which varied across

blocks). These workload measures included the DRT, TLX and a pseudo-TLX – similar to the Air Traffic Workload Index Task (Loft et al., 2015). The NASA-TLX is a well validated form of workload measurement administered via a short survey (around ten minutes) (Hart, 2006). The TLX encompasses key areas related to workload including temporal, mental and physical demand an individual experiences during a task. A full version of the task can be found at `www.keithv.com/software/nasatlx/nasatlx.html`. The TLX was administered at the end of the associated MOT block. The DRT used a tactile signal as in Innes, Evans, et al. (2020) which occurred across the duration of the block (including during the interrogation phase). The pseudo-TLX required participants to state a number from 0-10 specifying their perceived workload at the same frequency as the DRT. Results from the experiment followed expected trends; workload scores from the TLX increased with MOT difficulty, DRT response times increased with MOT difficulty and pseudo-TLX increased with MOT difficulty. Furthermore, MOT accuracy was unaffected by the type of workload measure. These trends provide further validation for the DRT-MOT framework, with convergent evidence highlighting that the DRT captured subjectively experienced workload trends induced by the MOT task.

In the study comparing the PVT, participants completed high and low difficulty blocks of MOT, interspersed with ten minute blocks of PVT. The PVT requires participants to respond to a stimulus - the presence of an increasing timer - at randomly distributed intervals between two and ten seconds. In this sense, the PVT is highly similar to the DRT as it requires participants to frequently detect and respond to a signal. The PVT however, intends to measure fatigue rather than workload. The only differences in assessing supposedly differing constructs comes in a slightly longer distribution of inter-trial intervals, and more importantly, in the absence, of a secondary task. The DRT is a dual-task design as opposed to the single task design of the PVT. Consequently, in this study, it is evident that the DRT and PVT are different, as the PVT has no simultaneous task – however, this was a key design feature. In keeping the PVT as a single task, I assessed whether cognitive workload effects (shown in DRT results) had an effect on fatigue, and vice-versa.

In the experiment, participants completed the DRT during the MOT tracking phase and, following blocks of MOT, participants completed ten minute blocks of PVT. Results from the DRT indicated a difference in workload between conditions of MOT difficulty, with low difficulty showing lower response times. The PVT did not capture this trend, or, more

so, the PVT was unaffected by the previous block of MOT. Furthermore, the DRT was unaffected by time across blocks – i.e. there was no difference in workload across time. The PVT did show some affects of time, with later blocks of PVT showing higher response times than early blocks. These results are important for the current thesis, as it shows that the DRT is measuring the construct of cognitive workload and is seemingly unaffected by fatigue effects – evidence for construct validity. Similarly, I provide evidence that the PVT is unaffected by any workload carryover from the preceding block of MOT.

Results from Experiment 1, and the supplemental validation experiments, provide evidence for the usefulness, and validity of the DRT-MOT paradigm. The following chapters extend on this, by using this paradigm to assess the quality of display information Chapter 4 and later to differentiate between groups, and be used as a selection metric Chapter 6. In the scope of this thesis, I view these upcoming chapters as the theoretical component of my research, where I take results from this chapter and extend this to answer real world problems.

Chapter 4

Application of the DRT as an evaluative tool

4.1 Adding Information - Helpful or Harmful?

As our attentional resource pool is limited (Kahneman, 1973), attending to too much information could be detrimental to performance. Take for example driving a new car, where the modern dashboard is more information rich than classic cars. This information is intended to be useful and helpful for the driver. However, attending to this information requires some cognitive resources and could potentially distract the driver from the main task – driving (Strayer et al., 2019; Thorpe et al., 2019). It is thus imperative that these stimuli be carefully evaluated to understand both the usefulness of the information and the cognitive demands or distraction it poses to the operator. It is intuitive that added information should aid performance, and this information is of greater value if it imposes low additional cognitive workload. Literature supports this notion, showing that useful information can improve performance (Eppler & Mengis, 2008; Vashitz, Shinar, & Blum, 2008), but only up to a certain point – essentially following an inverted–U hypothesis of performance (Hardman & Macchi, 2004). Under this hypothesis, where too much information is detrimental to performance, too little information, or information which is task irrelevant, can lead to similar performance detriments. This is similar to situational awareness literature, where low task engagement can have similar effects to highly demanding tasks. In this chapter, I test useful-by-design information using the DRT-MOT paradigm to measure the trade off between workload and performance when assistance is given in order to show that workload should be measured rather than assumed. The current chapter proposes a methodology of assessing the effects of increased information on task performance and cognitive workload – as well as the interaction between these factors.

Understanding the effects of increased information can be difficult, as this requires analysing both main task performance and latent variables related to cognitive demands. It is possible, and quite likely, that there is some task trade off between these factors - where adding more information may increase the demands on the operator, but may improve performance. This kind of trade off could asymptote or reach ceiling for performance and for demand, where after a certain point, information no longer increases performance or cognitive demands, and consequently is ignored. The amount of information given to the operator may also be a moderator of cognitive demands, where more useful information could reduce cognitive demand, or information may be distracting from the main task, and

therefore decrease the overall performance. Despite all of these possible explanations, it is clear that workload and performance need to be evaluated conjointly to understand the impact of information on performance and cognitive demands.

Measuring task performance is generally task dependent, and so observing the effects of increased information on task performance is straightforward in controlled environments. Understanding the impact of increased information on the latent variable of cognitive workload is more complex. Cognitive workload increases when more of our limited attentional resources are being used. With an individual performing multiple tasks, or tasks of greater difficulty, their cognitive workload increases with more mental resources exhausted (Kahneman, 1973; Strayer et al., 2013). With advances in computer technology, some research has shown general improvements in our multitasking performance (Haapalainen, Kim, Forlizzi, & Dey, 2010), and so adding additional information could aid the user without risk. Yet, there is a great amount of literature in driver distraction that shows this is not necessarily the case (Coleman et al., 2016; Strayer et al., 2017; Strayer & Johnston, 2001).

Chapters 2 and 3 have shown the DRT as an effective measure of cognitive workload, and further research has used the DRT to evaluate the effects of varied information in the driving environment, such as in-vehicle information systems (Strayer et al., 2019, 2017; Strayer, Turrill, et al., 2015). The DRT has been predominantly used in driver distraction literature to show the effects of distractions, such as mobile phones and conversations with passengers, on cognitive workload. Strayer et al. (2019, 2017) have shown the impacts of in-vehicle information systems, such as CarPlay and Android Auto, and smart assistants, from Apple and Google, on driver workload. This type of hands-free technology, which allows users to keep their eyes on the road and hands on the wheel, seems useful without impacting the driver. However, studies using the DRT have shown that workload significantly increased when interacting with these technologies - similar to when using a handheld mobile phone (Strayer et al., 2017).

Similarly, Haapalainen et al. (2010) used several measures including heart rate monitoring, eye tracking and subjective ratings to evaluate the cognitive load faced by users in gaming systems, finding that load increased under situations of greater multitasking. Thorpe et al. (2019) extend on this noting the importance of evaluating the impact of a

system and interface on cognitive workload. Further regarding displays and additional information, research such as Mayhew (1999) propose that user interface design should have a set of usability goals to meet, and this should include an assessment of cognitive workload, to show situations where adding information is no longer helpful. Understanding this trade off can be difficult, and so it is often equally difficult to find a point where information is no longer useful. This can often depend on the task outcomes - for example if the main objective of a task is performance with no need to consider cognitive workload, then increased amounts of cognitively demanding information may be added to a design. However, in an environment such as driving, it is imperative that drivers are not cognitively overloaded, and so information needs to be kept to a minimum - or at least to a standard which it is easily perceptible and intuitive, but not distracting. These are questions dependent on task and design, however, there is a method of concurrently assessing these factors.

Here I proposed a paradigm that included additional informative stimuli, which was developed to help the user's MOT performance, to evaluate both performance *and* workload factors. It is clear that "useful" additional information enables increased performance without impacting cognitive workload greatly; whereas "poor" additional information has no performance benefit (or even is detrimental to performance) or increases cognitive workload (or both). The current study used the DRT-MOT as used by Innes, Evans, et al. (2020). The additional information, which was termed *assistance*, was "attached" to MOT stimuli. The assistance was intended to be useful and usable, whilst also posing potential distraction or cognitive demand. There were two types of assistance used which were intended to differ in the type of processing demands required – one requiring greater visual distraction and demand, while the other imposed a memory cost. These types of stimuli were selected as distinct levels of cognitive processes, where the visual stimuli was easily processed and intuitive to use, whereas the text assistance required greater processing to encode stimuli. Both types of assistance consequently required either the same (for the visual) or different (for text) cognitive processes to the main task and both could be distracting for the participant, either by drawing their attention away from a focus point (visual) or due to extra cognitive resource costs (text). I evaluated both MOT performance and cognitive workload (as given by DRT results).

I conducted two experiments – a between-subjects paradigm and within-subjects

paradigm. Both experiments manipulated the assistance included in the MOT. The between-subjects paradigm manipulated assistance types between participants. In the within-subjects design, assistance was manipulated within participants. It was hypothesized that MOT accuracy would increase in conditions of added assistance, provided the assistance is useful. Secondly, it was hypothesized that DRT response times would increase with added assistance due to the additional distracting effects of the assisted stimuli.

4.2 Methods

Two experiments were undertaken: Experiment 2A was a between subjects experiment and Experiment 2B was a within subjects design. Both experiments used the same stimulus, as outlined in the Tasks section. The procedure differed slightly for the two designs. Rather than collecting more data for the between subjects design, a second experiment was conducted as the MOT procedure was adjusted to avoid potential ceiling effects. This is discussed below in Experiment 2B.

4.2.1 Participants & Design

Participants in Experiment 2A were 121 psychology students from the University of Newcastle, recruited online and reimbursed with course credit. The design was a 2x3 mix, with the within subject condition of assistance; assistance absent or assistance present; and the between subject condition based on the three types of assistance presented; reappearing target (64 participants), labels (53 participants) and both (53 participants). Each participant could opt to complete more than one condition, however, only 2 participants did.

Participants in Experiment 2B were 38 psychology students from the University of Newcastle, recruited online and reimbursed with course credit. The design was a three way within subjects design, with three levels of assistance; assistance absent (none), added labels (text) or reappearing target (reappear).

In both designs, only one MOT difficulty was interrogated - Experiment 2A used 4 dots to tack, and Experiment 2B used 3 dots to track. Exclusion criteria from section 3.2.4

was followed. From this, 49 data sets were removed for Experiment 2A, and 4 were removed in Experiment 2B.

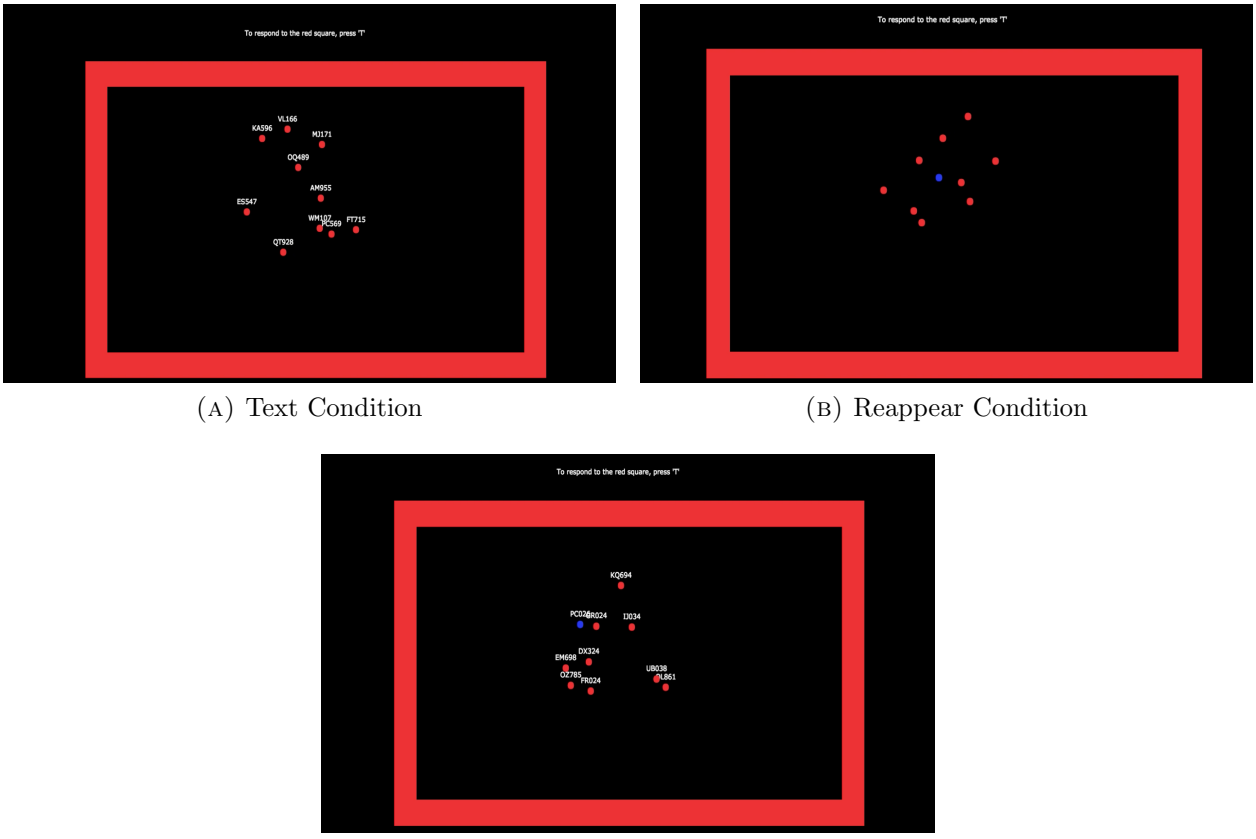
4.2.2 Tasks

4.2.2.1 MOT

The MOT involved participants tracking a number of 'target' dots within a display of distractor dots and followed the procedure of Experiment 2 in Innes, Evans, et al. (2020).

During the tracking phase, assistance could be added. The assistance was intended to be useful for the participants to identify the target dots. There were three types of added information; reappearing target, added labels and a third condition where both were available. There was also a condition where assistance was absent – the same as in Innes, Evans, et al. (2020). In the reappearing target condition, target dots would intermittently reappear (for one second) in blue during the tracking phase creating a kind of "reveal" effect. This added information was intended to help the participant confirm (or re-idenitfy) the target dots to track. Only one target dot was shown in blue at any given time during the tracking phase. The target dot to reappear was random, so the same target could reappear consecutively. The reappearance lasted for one second, before the next randomly selected target dot reappeared. This condition was intended to be useful – as participants are reminded of the target dot, but could be potentially distracting, as it may draw their attention from another dot or a focal point.

In the 'text' condition, randomized labels were added above all of the objects. These labels were five characters in length (two capital letters and three numbers). The labels remained above the dots throughout the encoding and tracking phase, but were removed for the interrogation phase. These labels were designed to resemble flight numbers on a flight controller display. They were intended to be useful as participants could quickly memorise the labels attached to the target dots. The information could also be distracting, as processing and memorising the labels may take too long, or could pose a significant demand on workload. The 'both' condition included both reappearing targets and labels. Examples of the three information conditions are shown in figure 4.1



(A) Text Condition

(B) Reappear Condition

(C) Both Condition

FIGURE 4.1: Examples of each added assistance condition in the design.

In Experiment 2A, participants completed blocks where no information was added and blocks where only one type of information (text, reappear or both) were added. For all blocks of MOT in Experiment 2A, participants were asked to track four target dots.

In Experiment 2B, participants completed blocks of the reappear condition, the text condition and assistance absent condition. In Experiment 2B, participants were asked to track three target dots.

4.2.2.2 DRT

The DRT closely followed the ISO standard (ISO:17488, 2016). Participants were asked to respond to the presence of a salient signal throughout the encoding and tracking phases of the MOT. The DRT signal was a red frame which bordered the MOT display (as shown in Figure 4.1). The signal onset was randomly distributed to occur every 3-5 seconds. Participants were asked to respond as fast as possible to the signal, which remained

onscreen for one second, or until a response was made (whichever occurred first). Participants responded via the keyboard (either "T" or "Y" keys, depending on handedness).

4.2.3 Procedure

Participants in both manipulations completed the task online. Participants were first given instructions on screen which introduced the DRT procedure and then were shown instructions regarding the MOT procedure. Following the instructions, participants completed a practice block of three MOT trials, followed by the test blocks (six blocks for Experiment 2A, nine for Experiment 2B).

Participants in Experiment 2A completed a total of six experimental blocks which alternated the presence of assistance (three blocks per condition). The types of assistance was randomised across participants. Each block of MOT consisted of seven trials, giving a total of 21 MOT trials for the condition with and without assistance. Within each trial, participants made five decisions for the MOT task. Participants were given breaks between blocks with the whole experiment taking around 30 minutes to complete.

Participants in Experiment 2B completed a total of nine experimental blocks which alternated the assistance given (three blocks per condition). There were 10 trials per block, giving a total of 30 MOT trials per condition. The sequential order of assistance presentation was randomised across participants. Participants were given breaks between blocks, with the whole experiment lasting for around 60 minutes.

Response time and accuracy were recorded for the MOT. Response time and proportion of misses were recorded in the DRT.

4.3 Results

For an overview of the analysis, see Chapter 3.2.4.

4.3.1 Experiment 2A

Experiment 2A was treated as a two-way mixed design, with the within-subject variable of presence of information (added assistance vs no assistance) and the between-subject variable of type of assistance (reappear, text or both). I assessed the response time and the proportion correct for responses to the MOT, as well as the RT and miss proportion for responses to the DRT. I used two-way Bayesian ANOVAs for each of the above measures of interest. For the ANOVA results, I will again refer to $BF_{inclusion}$ as the amount of evidence that data are likely under a model containing a given predictor compared to models without this predictor. Results of the Bayesian ANOVAs are presented in Table 4.1.

$BF_{inclusion}$	DRT RT	DRT Miss	MOT RT	MOT acc
Presence of Assistance	**> 1000**	**> 1000**	**551.11**	**> 1000**
Type of Assistance	2.75	49/100	**790.83**	**>1000**
Type:Presence	**8.82**	71/100	**438.91**	**>1000**

TABLE 4.1: $BF_{inclusion}$ factors across dependent variables (columns) for each predictor (rows). $BF_{inclusion}$ with sound, or greater, evidence are shown in bold. $BF_{inclusion}$ shown as fractions represent evidence for null effects of the given predictor. $BF_{inclusion}$ greater than three represent evidence for the effects of a given predictor, whilst $BF_{inclusion}$ less than a third represent evidence against the effects of a given predictor.

Overall, 121 data sets were used for analysis (49 reappear, 38 text, 34 both). The mean RT for the MOT was 0.806 s (SD = .22) and the mean accuracy in the MOT was 62.38% (SD = 12.04%). Figure 4.2 shows the change in performance from conditions of no assistance to adding assistance across the three groups. Mean accuracy appears to increase for the reappear and both conditions of added information, but declined for the text condition. The same trend was shown for MOT response time, with response time slowing in the reappear and both conditions when assistance was added, but becoming faster in the text condition. Bayesian ANOVAs confirmed these trends (Table 4.1), which indicated strong evidence for the effects of presence of assistance, type of assistance and the interaction between these. These results indicate that the presence of assistance did have an effect on MOT results (both accuracy and RT), but these varied across types of information. This is especially evident from the interaction effect, where the text group shows opposite trends to the other groups. Bayesian t-tests confirmed the reliability of these trends, showing evidence for a null difference between groups in assistance absent (none) condition for MOT response time (reappear vs text; $BF_{01} = 4.98$, reappear vs both; $BF_{01} = 4.57$, text vs both;

$BF_{01} = 4.63$) and most levels of MOT accuracy (reappear vs text; $BF_{01} = 4.25$, reappear vs both; $BF_{01} = 1.55$, text vs both; $BF_{01} = 3.23$). In conditions of added assistance, there was evidence shown for a null difference in MOT response time between the reappear and both conditions ($BF_{01} = 4.47$), with ambiguous evidence shown between the other conditions (reappear vs text; $BF_{10} = 0.74$, text vs both; $BF_{10} = 1.17$). For MOT accuracy in the assistance conditions, evidence was shown for a difference between reappear and text conditions ($BF_{10} > 1000$), with ambiguous evidence shown for a difference between other conditions (reappear vs both; $BF_{10} = 2.19$, text vs both; $BF_{10} = 2.06$)

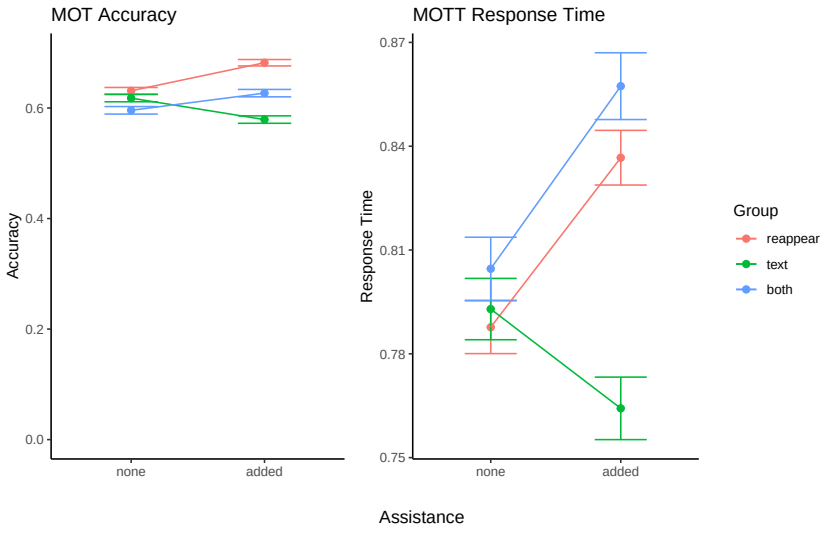


FIGURE 4.2: Performance on the MOT across groups. *Left Panel*: Accuracy in the MOT across modes of assistance (none or added) for the three groups. *Right Panel*: Mean response time in the MOT decision phase across modes of assistance for the three groups. Error bars shown are standard error.

Results from the DRT showed similar trends, as shown in Figure 4.3, with sound performance observed across participants and conditions for both mean DRT response time ($M = .55$, $SD = .14$) and DRT miss proportions ($M = .09$, $SD = .15$). The change in performance across groups and conditions for DRT metrics is observable in Figure 4.3, with DRT response time clearly affected by the interaction of type of assistance (group) and the presence of assistance, with the reappear and both conditions showing steeper slowing of response time to the DRT in the presence of assistance. Presence of assistance also appeared to have an effect on DRT miss proportion, with miss proportion increasing for the DRT when assistance was added. Results shown in Table 4.1 from the Bayesian ANOVA confirm these trends, with evidence for the inclusion of the effects of presence of assistance and the

interaction effect (presence of assistance and type of assistance) on DRT response times. For DRT miss proportions, evidence was only found for the effect of the presence of assistance, suggesting that adding assistance led to greater miss proportions.

Post-hoc Bayesian t-tests indicated that there was no difference in DRT response time between types of information (group) when assistance was *absent* (reappear vs text; $BF_{01} = 4.82$, reappear vs both; $BF_{01} = 4.93$, text vs both; $BF_{01} = 4.57$). In the *presence* of assistance, there was evidence for no difference between the reappear and both conditions ($BF_{01} = 4.63$), and ambiguous evidence for a difference between these conditions and the text condition.

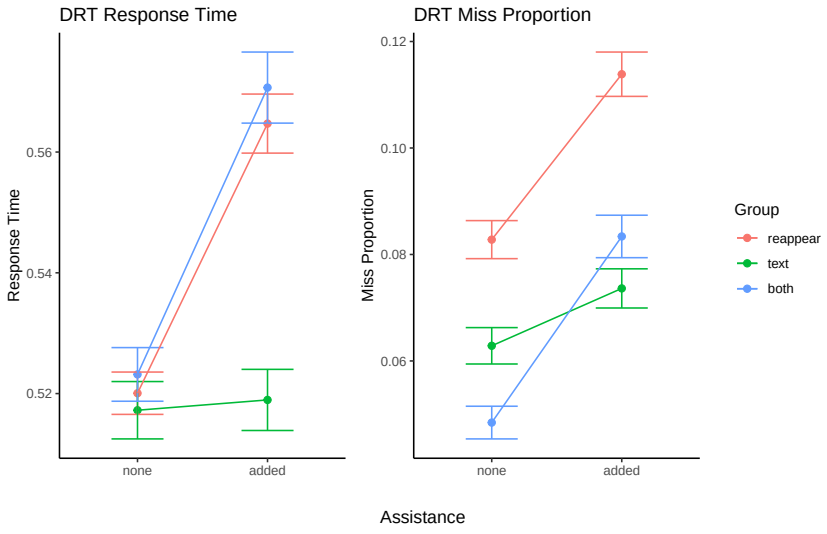


FIGURE 4.3: Performance on the DRT across groups. *Left Panel*: Mean response time to the DRT across modes of assistance (none or added) for the three groups. *Right Panel*: Mean miss proportion in the DRT across modes of assistance for the three groups. Error bars shown are standard error.

4.3.1.1 Individual Analysis

Finally, included in Figure 4.4 is an analysis of differences between individual participants across the six conditions, with the between subjects variable of information type shown in the columns, and the within subjects variable of presence of information presented across the rows. The individual analysis allows us to observe the differences between groups for the design. Clearly, the text condition fared the worst with the assistance added. It may be the case that participants ignored the label assistance or the labels interfered with tracking. However, there are participants within this group who have higher MOT accuracy in

the presence of the textual assistance. These individuals may have found a way to use this assistance more effectively. This highlights the importance of individual differences when investigating the usefulness of adding assistance in designs, as although the majority found the assistance detrimental, some participants were able to use the assistance as intended which aided performance.

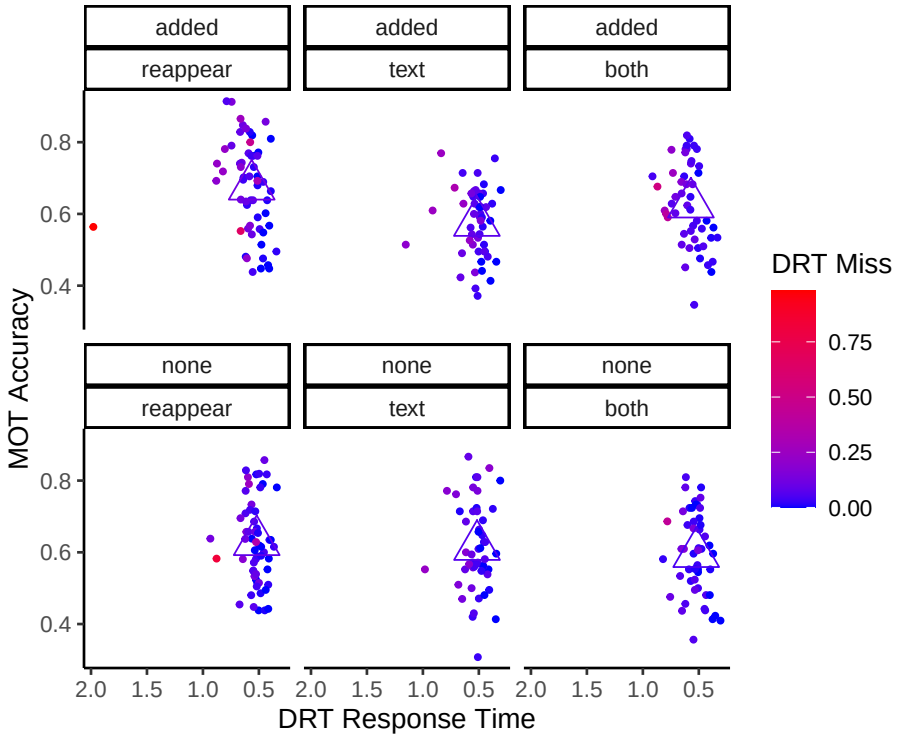


FIGURE 4.4: Individual performance for Experiment 2A, shown across conditions – information added shown as the top row and conditions shown by columns. MOT accuracy is shown on the y-axis and DRT response time is shown on the x-axis (note that DRT response time is shown descending from left to right). The colour of the dots indicates the number of misses, with blue being lower. The triangle shows the mean performance for each condition. Ideal performance would tend towards the top right of the figure with a blue shading.

4.3.2 Experiment 2B

Experiment 2B was treated as a three-way design, with varying types of added assistance (none, reappear or text) included as the within-subjects variable. Experiment 2B mainly varied from Experiment 2A in that difficulty of the MOT was reduced to avoid ceiling effects and participants completed two added assistance conditions as well as the control condition. After exclusion criteria, there were 34 data sets used in the analysis. The

mean MOT accuracy was 73.8% ($SD = 12.32$) and MOT response time was .99 s ($SD = .72$), whilst mean DRT response time was .55 s ($SD = .11$) and DRT miss proportion was 8.3% ($SD = 8.3$). Bayesian ANOVAs were conducted for the each of the four dependent variables (MOT accuracy and response time, DRT response time and miss proportion), with results shown in Table 4.2. Figure 4.5 and Figure 4.6 show the mean results across participants for dependent variables of the MOT and DRT respectively. It is worth noting that these figures differ from those above, as assistance conditions are shown on the x-axis.

Results from the MOT are shown in Figure 4.5, where accuracy seems to increase in the reappear assistance condition, but decrease for the text condition. For MOT results, similar to Experiment 2A, response time appeared to increase for the reappear condition. Bayesian ANOVAs confirmed these trends for accuracy, as seen in Table 4.2, however, evidence was shown for no effect of assistance type on MOT response time. Bayesian t-tests highlighted the reliability of this result, with strong evidence shown for a difference in MOT accuracy between the reappear condition and the two other conditions (reappear vs text; $BF_{10} > 1000$, reappear vs none; $BF_{10} > 1000$), and ambiguous evidence for a difference between the text and assistance absent conditions ($BF_{10} = 0.71$). These results indicate that MOT response times were unaffected by assistance condition, meaning that strategy was unlikely to differ for participants between conditions. However, the reappearance assistance appeared to aid participant performance, highlighting the usefulness of this type of assistance in the current design, whilst the text assistance appears to add no performance benefit. In fact, in conjunction with results from Experiment 2A, trends indicated that the text condition not only fails to assist participants, but may actually decrease MOT performance.

$BF_{inclusion}$	DRT RT	DRT Miss	MOT RT	MOT Acc
Type of Assistance	**>1000**	**7.38**	**1/9.1**	**>1000**

TABLE 4.2: $BF_{inclusion}$ factors across dependent variables (columns) for the predictor (type of assistance). $BF_{inclusion}$ with sound, or greater, evidence are shown in bold.

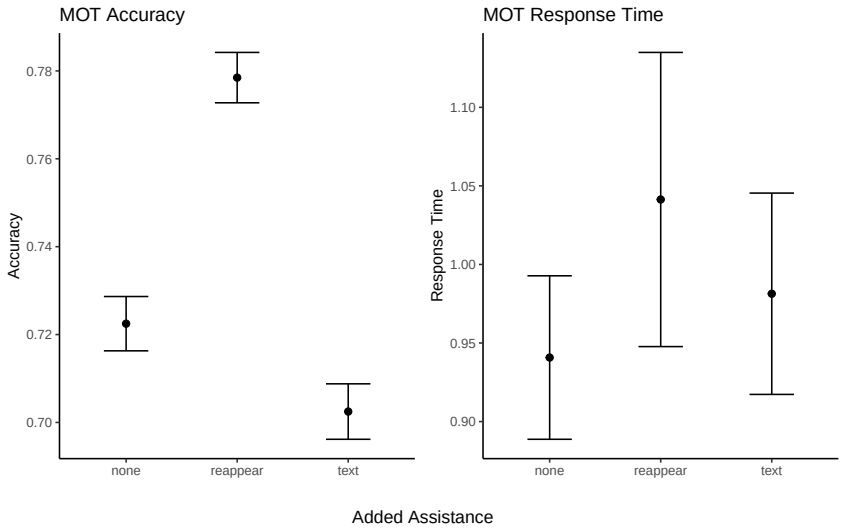


FIGURE 4.5: Performance on the MOT across assistance conditions. *Left Panel*: Accuracy in the MOT across types of assistance. *Right Panel*: Mean response time in the MOT decision phase across types of assistance. Error bars shown are standard error.

DRT results, as in Figure 4.6, show related trends to the MOT results. In the DRT, the reappear assistance condition appears to have the highest associated response times, whereas the text assistance condition appears to have no effect on DRT response time. This is different in DRT miss proportions, where it is evident that the greatest lapse proportion is observed in the text assistance condition, with the reappear assistance condition also showing increase miss proportions. Bayesian ANOVAs confirmed these trends for DRT response time, as seen in Table 4.2, with strong evidence for an effect of assistance type. Bayesian t-tests confirmed these trends, with the reappear assistance condition showing strong evidence for differences from the other assistance conditions (reappear vs text; $BF_{10} = 70.18$, reappear vs none; $BF_{10} = 231.42$). Ambiguous evidence was shown for a difference between the assistance absent condition (none) and the text condition for mean DRT response time ($BF_{10} = 2.18$). Furthermore, a Bayesian ANOVA showed evidence for a difference in DRT miss proportions across conditions. Bayesian t-tests highlighted evidence for a difference between the text condition and the assistance absent condition ($BF_{10} = 18.95$), however, there was ambiguous evidence for differences between other conditions. The DRT results present a telling story, as response times are seemingly unaffected in the text condition (compared to the absence of information condition), however, participants show an increase in lapses – a result which indicates higher cognitive workload. Furthermore, the reappear condition similarly shows an increase in participants cognitive workload, highlighting that both assistance conditions

were associated with a greater allocation of mental resources. Alternatively, the link between DRT response time and MOT accuracy may be the result of a speed accuracy trade-off, as the reappear condition shows the highest MOT accuracy and slowest DRT response time. This may also represent a link between cognitive workload and the speed-accuracy trade-off phenomenon, where high workload leads to a more prevalent trade-off.

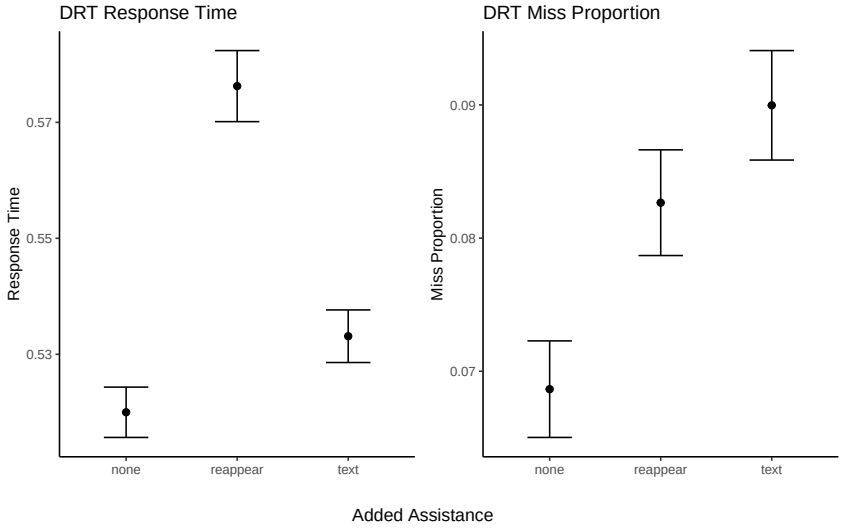


FIGURE 4.6: Performance on the MOT across assistance conditions. *Left Panel*: Accuracy in the MOT across types of assistance. *Right Panel*: Mean response time in the MOT decision phase across types of assistance. Error bars shown are standard error.

4.3.2.1 Individual Analysis

Akin to results of Experiment 2A, Figure 4.7 shows individual performance across the assistance conditions. Results similarly show that some individuals are able to use the text condition effectively, to attain a higher accuracy, however, in Experiment 2B, there are far fewer of these participants. This indicates that rather than participants showing greater efficiency with textual assistance, these individuals were more efficient in the MOT task than their peers. Furthermore, as Experiment 2A used 4 dots to track, participants may have been at their performance ceiling, and so presence of assistance had no true impact on performance. In reducing the difficulty of Experiment 2B, results indicated that this may have been the case for individuals who showed higher performance with textual assistance in Experiment 2A, as there were a limited number of individuals who showed a preference for the text assistance (in MOT and DRT performance).

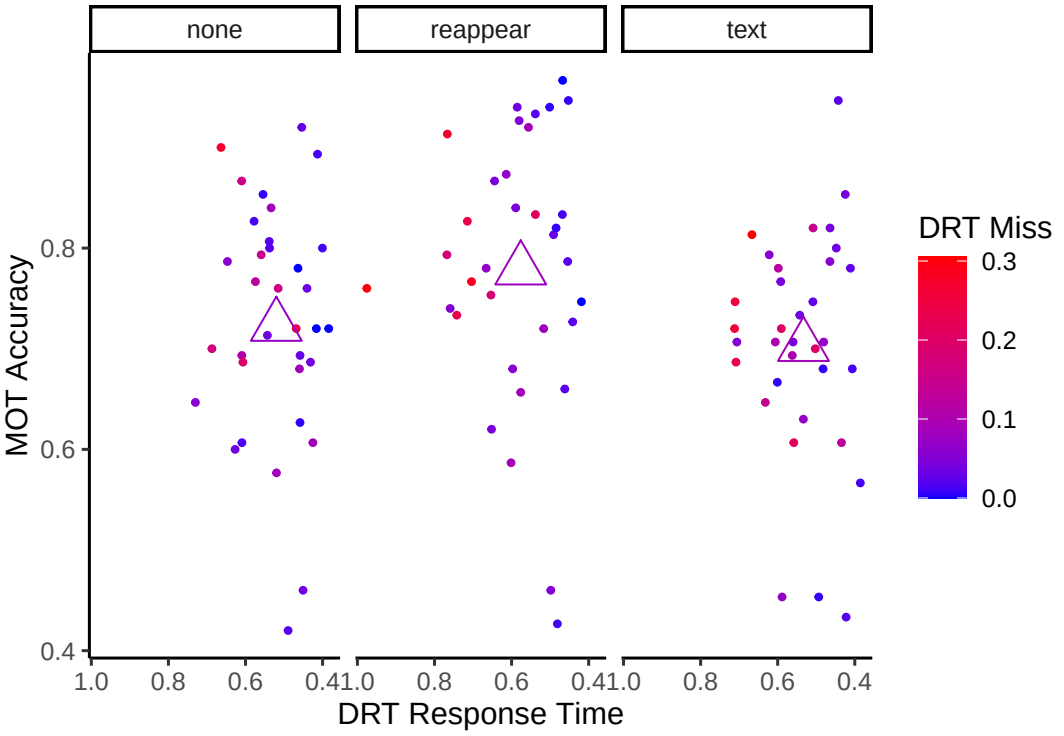


FIGURE 4.7

4.4 Discussion

The current study aimed to evaluate the effects of additional "useful" information on cognitive workload. Both the within and between subjects studies showed highly comparable results across the different types of information added. For Experiment 2A, it was clear that the MOT task was difficult, and so the difficulty was reduced for Experiment 2B to control for ceiling effects.

In both studies, it was evident that accuracy was affected by the type of assistance given to participants. When the assistance given was the reappearing target, accuracy increased. However, in the text condition, where labels were attached, participants' accuracy decreased. The condition which included both types of assistance (in Experiment 2A) showed an increase in accuracy, but a smaller increase than in the reappearing target only condition. These results provide ambiguous support of the hypothesis, with the reoccurring target location leading to an increase in accuracy, however, the label condition lead to a *decrease* in accuracy.

When evaluating cognitive workload in both studies, there is a clear trade off observed. In the reappearing target condition, DRT response time increased. In the text condition, there was no increase to response time. In the within subjects "both" condition, response times also increased. These results support the hypothesis that additional assistance leads to an increase in cognitive workload. Again however, the label condition did not support the hypothesis, with observed workload higher in this condition, with a *decrease* in MOT performance. Overall however, the trends remain across both experiments – adding "useful" assistance costs workload but corresponds with performance increases; and adding assistance which is not useful generally adds no workload, but has minimal performance benefits.

When combining these MOT and DRT results, a pattern becomes evident. In the target reoccurring condition, accuracy of tracking was improved, however cognitive workload increased. This provides evidence that the assistance was useful, but required greater cognitive resources to process the information. In the label condition, the opposite of this seems to be the case, with no increase to cognitive workload and a decline to accuracy of tracking. This type of assistance is detrimental to a design as it appears to be no more distracting (in terms of cognitive resources), but actually leads to a decline in performance. This could be due to participants ignoring the distracting information and directing more resources towards the DRT, consequently sacrificing performance in the MOT.

There were differences between experiments 2A and 2B, which are likely due to the difficulty difference in the MOT. Trends appeared to hold across the two difficulty conditions and between the between vs. within subjects manipulation. Experiment 2B (with three dots to track) showed similar overall mean DRT response times, but greater MOT accuracy. Interestingly, despite the clear differences, for all assistance conditions, in MOT accuracy between experiments (where Experiment 2B had higher accuracy), there was no difference across all assistance conditions for DRT response times. This finding is not subject to statistical analysis due to the groups completing alternate experiments, however, it does show the importance of analysing the trade-off between tasks. In Experiment 2A, participant workload was likely higher than participants in Experiment 2B, however, this difference was not captured by the DRT. Instead however, participants showed a greater trade-off with MOT performance, with participants in Experiment 2A performing closer to chance levels

than participants in Experiment 2B (who showed higher than chance accuracy). Two of the possible explanations for this outcome are that participants in Experiment 2A gave up on the MOT to focus on the DRT with the difficulty too high, or alternatively there is a speed-accuracy trade-off which is moderated by the difficulty.

It is also worth noting the individual differences observed. These are best viewed in the results of Experiment 2B, where we can directly compare conditions between individuals. In this Experiment, results showed that for 12 subjects, MOT accuracy was higher in the text condition than no assistance. Furthermore, results indicated that 7 participants showed higher accuracy in the text condition than the reappearance condition. These results highlight the individual differences that do exist, which should be accounted for in designing displays or evaluating additional assistance.

These results show two divergent examples of adding assistance into a display. An example of useful assistance was shown by the reappearing target condition. In this condition, performance was enhanced, however it came at a cost to workload as processing the assistance required a greater amount of mental resources. An example of poor assistance was shown in the label condition. Labels were intended to be useful as participants could dedicate resources to memorising labels attached to targets in order to make tracking easier. However, results showed that participants were using no extra resources and that the information actually led to a decline in performance. The reasons for this are not entirely clear. Some participants may have found the labels too difficult to memorize, consequently leading to a false memory trace when tracking. Alternatively, some participants may have opted to ignore the labels but then later been distracted by them. There are also design limitations, as the labels may have appeared too distant from the target object, or targets may not have remained highlighted for long enough for participants to scan all three or four and associate this with the target. When observing individual data, it is evident that some participants could actually use the information to their advantage, but, similar to the reoccurring target condition, suffered in the workload measure. Further research and analysis, including measures such as eye-tracking or subjective surveys are required to get a deeper insight into the strategy adopted in this text condition. At a surface level however, it is clear that the two sources of assistance provide sound examples of useful, costly assistance and ineffective, unnecessary assistance.

The current study links closely to user interface and design literature, with designers more frequently adopting technology into user interfaces and displays. This technology may seem intuitive and useful at a surface level, however, the necessity of the technology and the way in which it is inputted need to be evaluated thoroughly to form a holistic view of the impacts. Take for example the reappearing target condition - in a display which is already highly mentally taxing, adding this information may be useful for the operators performance, however could lead to a potential cognitive overload scenario as the stimulus requires deeper processing. If an operator is already highly loaded, this information may be more harmful than helpful. In another display where workload demands were low and the importance was placed on performance, this type of assistance would be highly useful. Mayhew (1999) propose that designers should establish a set of usability goals that an interface should achieve, and the current study shows the importance of workload evaluation as a part of those usability goals. As outlined above, user interface literature has noted the importance of workload evaluation (Dan & Reiner, 2017; Gerjets, Walter, Rosenstiel, Bogdan, & Zander, 2014), and the current study provides useful examples of the perils and pitfalls of adding assistance into a display.

It is further important to consider the context of such research, with different environments posing unique challenges and requirements. As discussed earlier, different environments may place greater importance on one factor over another. For example, in driving, there is a greater amount of importance placed on cognitive workload than task performance as task performance is relatively indistinguishable above a threshold (i.e. most driving is considered "good" performance unless a serious error is made). However, if the driver is distracted, it could lead to serious errors or lapses in concentration, and therefore it is imperative that workload remains at a minimum. In other environments, task performance may be preferred over workload, as cognitive overload may not have serious consequences. These factors must first be considered in any design as they are important indicators of the overall success of added assistance. Future studies could potentially evaluate this by giving unique instructions across conditions to place importance on performance or limiting cognitive load - similar to studies of speed-accuracy trade off.

4.4.1 Limitations and Future Directions

The current study was limited by several design elements which could influence results. Initially, the four dot condition of the MOT in the between subjects study may have been too difficult and consequently we were observing ceiling effects on cognitive workload in some conditions of assistance. This was rectified in the within subjects experiment, and seemed to have an impact with more clear differences shown between assistance conditions. Secondly, the trade off between workload and performance was difficult to directly associate with one factor, especially in the text condition, as parts of the design may have distracted participants in both tasks rather than just one. This is a key tenet of the DRT, however, if *strategy* or task preference was important to analysis, this is a limitation. Finally, in the text condition, the labels were turned off for the interrogation phase of the MOT. This could have caused a distraction or could have affected performance if participants were tracking the labels and not the associated dot. Future studies could place the label closer to the associated dot, or overlay the label on the target. Alternatively, future studies could keep the labels active during the interrogation to assess whether the participants are able to use this assistance. Other factors such as increasing the encoding time could also be considered in future designs.

Chapter 5

A novel practical application of the DRT

The current chapter shows a practical application of the DRT methodology discussed in Chapter 4. The contents of this chapter are taken from Innes, Howard, et al. (2020) which is published in the Journal of the Human Factors and Ergonomics society[1]. The paper was completed as part of this thesis, and fits closely within the scope of the thesis as a further application of cognitive workload measurement, in a novel environment. Similar to Chapter 4, I use the DRT to evaluate the impact of different types of information – however, in this paper, I evaluate the impact of heads-up display information on the workload and performance of helicopter pilots. My role on this paper as the main author was to develop methodology, collect data, conduct analysis and writing. The experiment was conducted in collaboration with Airbus Defense and Space and Hensoldt Sensor Systems.[2]

5.1 Measuring Workload in Aviation

With more information present and technological advances, our ability to multitask has seemingly improved (Haapalainen et al., 2010), however, there is substantial literature on driver distraction and cognitive workload that suggest this is not the case (Strayer et al., 2017; Strayer & Johnston, 2001; Strayer, Turrill, et al., 2015). Both added visual stimuli and seemingly useful information systems can lead to detrimental distraction due to cognitive load in drivers (Lee et al., 2008; Strayer et al., 2017). Here, we offer a novel and unique workload-capacity assessment of helicopter pilots. Specifically, technological advances enable rich information to be projected into pilots' heads-up displays (HUDs), but the impact of this extra information on cognitive demand is not well understood. Here we ask; can too much information be detrimental to performance? To answer this question, we tested highly qualified helicopter pilots in a flight simulator in varying environmental and HUD settings.

Cognitive demands and distractions are difficult to assess within a multitasking environment. Adding to the number of items to process (or increasing the difficulty of these items to process) causes a greater depletion of limited attentional resources (Kahneman, 1973; Townsend & Eidels, 2011). When attentional resources are low, responses are impaired and

[1]As this paper is published elsewhere, I have included footnotes where formatting or reporting is inconsistent and where content may be repetitive. The paper can be found at `https://doi.org/10.1177/0018720820945409`

[2]Much of the introduction below is covered in earlier chapters. Here I show how these principles apply to real world aviation settings. This content may seem repetitive from earlier chapters.

we experience a diminished ability to process and react to the demands at hand. Such is the case when completing cognitive tasks while driving – our performance is diminished in both tasks (Watson & Strayer, 2010). Here we define cognitive workload as the level of cognitive demand placed on an individual from a task/s and distraction as scenarios where the individuals attention is drawn away from the main task/s. (Lee et al., 2008)

The detection response task (DRT) adds an additional task that measures residual resources via a simple detection task. In the DRT, which is a standardised procedure (ISO:17488, 2016), participants are asked to respond as quickly as possible to a salient stimuli, which is administered frequently, whilst performing another task. Longer response times and increased misses correspond to higher cognitive workload (Strayer et al., 2013). Reactions are impaired when people are subjected to greater task demands, leaving fewer resources to allocate to the DRT. As an example, Strayer et al. (2013) showed that DRT response times for car drivers increased with the presence of a passenger or when talking on a mobile phone (both forms of distraction), similar to the increase when performing an operation span task. The sources of cognitive load mentioned above are external to the task at hand—it is not necessary to talk on the phone while driving—but systems related to completing the task, such as a user interface, can also impose cognitive workload. In extreme cases, a user interface can undermine its intended purpose of assisting the user by presenting too much information or interrupting relevant tasks (Johnson & Wiles, 2003).

User interfaces and other information delivery systems should therefore present only as much information as a user needs in an unobtrusive way (Haapalainen et al., 2010). A complication for user interface developers is that the amount of information a user needs may change as the user's workload state changes — a level of information that may be appropriate in one context may overload the user in another. A solution to this issue is to change the amount or presentation of information in real time, based on the user's cognitive capacity. A concurrent measure of workload is one necessary step in developing these *adaptive interfaces*.

A large body of cognitive workload research is centred around distraction in driving environments, yet this research is equally critical to the understanding of human-machine interactions in aviation. Helicopter and aeroplane cockpits are both extremely demanding environments, with a plethora of interfaces delivering multiple streams of information concerning air-speed, heading, fuel, obstacles and alike. C. D. Wickens (2002b) outlines

interlinking factors crucial to human interaction with aircraft, and highlights that much psychological research related to these factors has been conducted in isolation. Further, Kantowitz and Casper (2017) reference the increasing amount of technology and automation in aviation, which impacts crew workload – noting that studies of attention may assist in solving workload related problems in aviation environments. As distracted driving literature has shown, understanding the impact of this technology is vital, with the literature informing policy and technological development (Strayer, Cooper, Turrill, Coleman, & Hopman, 2015; Young et al., 2013). In aviation, Huttunen et al. (2011) and Hannula, Huttunen, Koskelo, Laitinen, and Leino (2008) both evaluated cognitive workload using the speech prosody and psychopsyhiological stress (PPS) indicators respectively. Whilst these measures are effective in assessing their related constructs, they may not be reliable indicators when assessing workload induced by technological factors. Previous work by Zimmermann et al. (2019) also aimed to assess the usefulness of additional HUD information, in the helicopter setting, with findings indicating that pilots flew more effectively under conditions of more information. Further, in the military setting where this technology is most used, landings are far more frequent and difficult, meaning that the HUD information allows a safer environment in critical scenarios. However, the measure of cognitive workload used in this study – the NASA Task Load Index – provided inconclusive results regarding cognitive workload. Evidently, with technology and automation constantly developing in avionics, literature stresses the need for evaluation of workload to ensure usability of such technology.

Some pilots and avionics developers operate as though more available information can only be beneficial, but this overlooks cognitive workload factors (Thorpe et al., 2019). Inversely, the type of information given to pilots may reduce their workload if the information is more readily perceived and easily processed, such as information which is 3D and more naturalistic (Dan & Reiner, 2017; Gerjets et al., 2014) [3]. In the current study we use the DRT to assess the workload demands arising from changes to the environment and the way information is presented (referred to as level of symbology). As the DRT assesses cognitive workload through residual capacity, we expect results from the DRT to translate from distracted driving literature to aviation environments.

[3]In linking this thread with the objectives of this thesis, it is clear that designers require the tools to evaluate this trade off between what is plausible and subjectively acceptable, compared to what objective measures actually indicate in cognitive workload.

The purpose of the current study was to evaluate the effectiveness and sensitivity of the DRT in a helicopter simulator environment, by varying the difficulty (environmental factors) of simulated flight conditions. The helicopter flight task was completed in a high-fidelity flight simulator. Flight simulators are widely used and well validated training facilities (Hays, Jacobs, Prince, & Salas, 1992; Roenker, Cissell, Ball, Wadley, & Edwards, 2003), so evaluation of cognitive workload in a simulator could facilitate deeper understanding of pilots' cognitive demands. Further, we used the DRT as a tool to measure the impact of added visual information ("symbology") in a HUD on helicopter pilots cognitive workload. We compared industry standard HUD symbology to new, more information rich symbology[4], as well as a control condition with no symbology. For a full overview of the technology input to the HUD see Zimmermann et al. (2019). Despite the limited number of participants, we placed a high importance on ecological validity of the task, designing a flight path that emphasised a realistic scenario, and testing pilots who were highly familiar with military helicopter environments.

It was expected that more information given to pilots would result in better flight outcomes. However, it was also anticipated that more information would lead to an increase in cognitive workload, similar to results reported by Strayer, Cooper, Turrill, Coleman, and Hopman (2016), Strayer, Turrill, et al. (2015) and Strayer et al. (2019). We first hypothesized that increased symbology would increase flight performance and landing accuracy, similar to results from Zimmermann et al. (2019). We also hypothesized that DRT response times would increase with lower visual acuity (i.e. worse simulated weather conditions). Finally, we hypothesized that DRT response times would increase with added symbology.

[4]Here, "information rich symbology" means that pilots were able to see a large amount of input within their HUD. Moving from basic symbology such as speed, altitude and flight lines to higher levels of symbology, such as 3D terrain grids, landing assistance and more. Examples of this can be seen in Appendix A (A).

5.2 Method

5.2.1 Participants

Eight pilots with experience in helicopter simulators and 2D symbology undertook the study. All pilots were male, had over 2,000 hours flying experience and extensive simulator experience. Seven pilots were recruited from the Airbus Brisbane facility, with one recruited from Hensoldt staff. Pilots recruited from the Airbus facility were either current military personnel or involved in testing or training. It was imperative that we tested highly trained and experienced personnel to ensure that confounding variables were limited; especially related to familiarity with the large-platform helicopter equipment and the advanced heads-up display. This research was approved by the Human Research Ethics Committee at the University of Newcastle (HREC-2013-0250).

5.2.2 Equipment

A helicopter simulator was used as the background during data collection. Data was collected in an Airbus MRH90 Taipan Multi Role helicopter simulator. The simulator incorporated three partially overlapping screens which made up 200° x 40° field of vision. The participant sat at a radius of approximately two metres from the screen. Controls in the simulator included a collective shaft, cyclic shaft and two foot pedals. The participants were shown an electronic map and a multi-function display, which indicated altitude, ground speed, collective power and helicopter roll. Participants were also fitted with a headpiece which was placed over the participants eyes. The headpiece acted as goggles, so that the participant could still see the simulator. In conditions where symbology was added, additional information was overlaid in their visual field. The location and angle of the headpiece was tracked at high rate so that information projected into the visual field mapped accurately and dynamically onto the visual environment.

A DRT device was used, closely adhering to ISO 17488 2016. The DRT device included a vibrating pad, which was taped to the participant's skin near their shoulder, and a response button, which was attached to the collective shaft nearest to where the pilots thumb

rested. Engström et al. (2013) provide evidence for the tactile DRT as a sensitive measure of cognitive workload, finding similar trends to the use of a visual stimulus. Furthermore, Cooper et al. (2016) suggest the tactile DRT is most effective for cutting down potential visual conflicts. With an already crowded visual environment, we proposed the use of the tactile DRT to limit visual workload effects.

5.2.3 Stimuli and Design

Each participant completed two simultaneous tasks – the flight simulation and the DRT. For the DRT, a short stimulus was elicited via a vibration. The participant was required to respond via the response button to each iteration of the stimulus. The stimulus lasted for one second (or until the response button was pressed, whichever came first). The DRT stimulus was elicited at an interval of 3 - 5 seconds and occurred for the duration of each simulated flight. Responses entered before the onset of the next vibration stimulus were deemed "hits", and failures to respond within 2.5 seconds were deemed a "miss". Second (and subsequent) responses entered before the onset of the next stimulus, as well as responses faster than 0.1 seconds, were deemed "false alarms". Response time was measured as the time between the onset of the vibration stimulus and the pressing of the switch.

The flight simulation involved participants undertaking a predetermined flight path with multiple objectives throughout. There were two conditions of visual environment: Day and Night. In all conditions, air traffic was absent, wind speed was set at 5km/h and weather was set to have no cloud or rain. The only parameters that varied were visibility (distance in meters), time of day, dust (on or off) and FLIR (on or off). The dust parameter related to simulated "brown-out", where simulated dust would inhibit pilots view below a certain altitude ($\sim$ 100 feet). FLIR (forward looking infrared radar) is an industry standard night vision technology, used only in the night conditions. A full summary of conditions can be seen in Figure 5.1.

We used three levels of HUD information; no symbology – where there was no information projected onto the pilots' HUD; 2D symbology – the generic two-dimensional information projected to the pilots' helmet (see Appendix A (A) for more details); or conformal 3D symbology – information which appears to be overlaid onto the simulated environment, as

well as the generic 2D information (see Appendix A (A) for more details; for a full overview of Hensoldt's Sferion assistance system, see Münsterer et al. (2014)). In the 2D condition, pilots were shown industry standard symbology which included speed, heading, altitude, geographic coordinates and distance to the LZ were displayed. Münsterer et al. (2014) provides a good example of standard 2D information. The 2D symbology condition was made as similar as possible to the standard heads-up display used by military helicopter pilots in modern large-platform helicopters.

In the 3D condition, symbology included the 2D information, horizon lines, ridge lines, landscape grids, highlighted obstacles and LZ virtual towers which assisted in guiding the pilot. Figure 15(d) in Zimmermann et al. (2019) and Figures 8–11 in Münsterer et al. (2014), provide good examples of the 3D symbology condition. The 3D symbology condition contained extra information, and the condition without symbology contained less. In the 3D symbology condition, all 2D symbology was shown, as well as the LIDAR (light detection and ranging laser sensor) information. This included a grid over the environment, contours, LZ information, horizon line and helicopter position. For an example of the three symbology conditions, see Appendix A (A).

In conditions without symbology, the headpiece remained fixed to the participants but displayed no visual information in the Day or Dust conditions. In the Night condition without symbology, FLIR (forward looking infrared radar) information was projected in the headpiece, with no additional symbology. In the 2D symbology condition, ground speed, radial altitude, location zone distance, and helicopter position were shown, as well as basic indicators for the waypoint and landing zone (LZ).

The study used a 2x3 factorial design, with two levels of visual environment (Day or Night) and two levels of Symbology (2D or 3D). Additionally, a condition without symbology was presented in either the Day or Night environmental condition was included. The visual environment in this condition was counterbalanced across pilots. Each pilot therefore completed five conditions – four with each level of symbology and visual environment, and one of two possible no-symbology conditions.

	NO SYMBOLOGY (0D) Headpiece off	**2D MINIMAL (2D)** *LIDAR off GND SPD, RAD ALT, LZ DST, LINE	**3D MAX SYMBOLOGY (3D)** *LIDAR on All symbology.
DAY VIS = 12000 TIME = 1600 DUST = OFF	DAY 0D	DAY 2D	DAY 3D
NIGHT VIS = 12000 TIME = 2000 DUST = OFF FLIR = (ON) FLIRTIME =2000 FLIRVIS= 2400	NIGHT 0D	NIGHT 2D	NIGHT 3D

FIGURE 5.1: Full table of experimental conditions. The table shows the 2 x 2 within subjects design with the added between-subjects conditions without symbology (shaded in grey). Each condition maintained strictly controlled simulator settings with the exception of those listed under the title. in the table "VIS" stands for the visual range (in metres); "TIME" indicates the hour of day in the simulator; "Dust" indicates whether brown out was on or off for landings; "FLIR" stands for Forward Looking Infrared Radar; "FLIRTIME" indicates the setting for FLIR time of day - a brightness setting; "FLIRVIS" indicates the visibility range setting for FLIR. Symbology conditions vary on the information which is displayed. In 0D, the headpiece is switched off (aside from FLIR on in the night condition). In 2D - Ground Speed, Radial Altitude, Landing Zones Distance and a horizon line are displayed. In 3D symbology, all of the 2D symbology with additional landing zone displays and LIDAR (Light detection and ranging) is displayed. For a further breakdown of the symbology conditions, see Appendix A (A). The shaded boxes show the two randomised between subjects conditions - pilots only completed one of these.

5.2.4 Procedure

All participants were familiar with the simulator environment, and were given instructions about the DRT. Participants were not instructed to preference either the DRT or the flight task, but were instructed that performance was measured across both[5]. The designated flight path was outlined to the participants. They were given several minutes of flight

[5]Despite this, pilots tended to give preference to the simulated flight due to training, where they are trained to "task shed" in difficult periods. The decision to not specify a preferred task here was motivated by the study by Conti et al. (2012)

time to adjust to the simulator before completing a practice run on the designated path. Following this, participants were given five practice DRT trials in isolation. Participants then began the experiment. The DRT commenced as soon as the pilot lifted the collective shaft for each condition.

The flight path was identical for all six conditions. The flight path took approximately 13 minutes to complete. Pilots were given verbal instructions during the flight to inform them of the objectives. Objectives included items such as passing a specified point at a target altitude and speed (known as "gates"), as well as specific landing scenarios, for example, landing in the centre of a sand bank. The flight path was divided into six sections, each with a different requirement, such as gates or landings. The objectives for the whole flight included two landings (one of which had poor visibility), an aborted landing, and three set "gates" to pass through at target speed and altitude. Furthermore, pilots were given directions on speed and altitude for each section, as well as specific navigation instructions. For a full breakdown of the flight path, see Appendix A (A).

Participants each completed five of the six conditions. The order in which the conditions were presented to participants was pseudo-randomised; the no-symbology condition was never presented first, to account for the pilots lack of familiarity with the flight path. If, during a flight trial, the participant crashed or there were any technical issues, the run was restarted. Responses in these trials were recorded separately. Participants were given breaks between flights. All flight data was recorded. DRT response times and misses were recorded. [6]

5.3 Results

In order to include the effect of the "no symbology" condition, we treated our study as a 2x3 design, with within-subject variables of time (day or night) and mixed variables

[6]Results in this section are not presented according to Section 3.2.4. Here BF_{10} is used to represent evidence for the winning ANOVA model, and $BF_{inclusion}$ is not shown. BF_{10} values are not shown as $BF_{10} > 1000$, but rather show exact estimates. Further, graphs appear in a different format. The design for this study was mixed, where the time of day variable was within subjects and the symbology was mixed (within subjects for 2D and 3D, but varied for the no symbology condition.

of symbology (none, 2D, 3D). Flight performance was given by a number of indicators selected after consultation with experts and aviation literature (Krueger, Armstrong, & Cisco, 1985). These indicators were measured objectively and were quantifiable, as well as remaining relevant to the task. Indicators assessed were landing data, in-flight targets and overall flight variability. The main reason to evaluate flight quality was to ensure there was no task trade-off between the flight and the DRT. DRT response time and misses were analysed. We removed 4 sets of flight data due to crashes. These crashes were generally simulator related, such as a failure to calibrate the headpiece within the environment. These flights could provide interesting insight into pilot behavior under load, however, results from this data were uninformative due to the lack of data and varying crash time points.

For the workload measure we assessed mean DRT response time and the proportion of lapses. For each metric we completed Bayesian ANOVAs for the environmental conditions, symbology conditions and the interaction. All analysis was completed using the statistical program JASP (JASP Team, 2019) using default priors. Bayesian ANOVAs operate in much the same way as traditional frequentist ANOVAs, but with a key advantage: Bayesian ANOVAs can separately identify evidence in favor of an effect vs. evidence in favor of no effect (i.e. positive evidence for the null hypothesis). This is communicated through Bayes factors (BFs), which compare the likelihood of the null hypothesis (H_0) – which assumes no difference between conditions – against the likelihood of the alternative hypotheses (H_1) – which assumes a difference between conditions. Bayesian inference has become a standard approach in many fields because of its advantages over frequentist methods (Wagenmakers, Lee, Lodewyckx, & Iverson, 2008). For clarity, we report all BFs in the direction of showing evidence in favor of the alternative hypotheses (BF_{10}). This means that larger BFs indicate more evidence *for* a difference between conditions. BFs near 1 indicate ambiguous evidence – the likelihood of the null and alternative hypotheses are about equal – and BFs much smaller than one indicate evidence in favor of the null hypothesis. We referred to (Jeffreys, 1961) for interpretation of BF_{10}.

5.3.1 Flight Metrics

We assessed the accuracy of landing data by borrowing appropriate precision measures from ballistic sciences. Participants were instructed to land at a specified and marked point in the virtual environment (centre of a sand bank). We measured the absolute distance from this landing zone (LZ) to the actual landing location ("landing error") and the "circular error probable" (CEP), which is the median error radius (Nelson, 1988, p.1). We analysed landing data using CEP for each the first landing zone (LZ1) and third (LZ3). Landings at LZ2 were aborted – by design. LZ3 did not utilize any landing symbology, making it a useful control condition. At LZ1, landing accuracy (defined by median absolute distance from the defined LZ in meters) was significantly improved with 3D landing symbology. The average distance from the defined LZ was 6m (SD = 6m) with 3D symbology, compared with 40m (SD = 41m) for conditions without symbology, and 61m (SD = 65m) with 2D Symbology. A Bayesian repeated measures ANOVA showed a reliable main effect of symbology ($BF_{10} = 3.01$), although evidence was ambiguous for the difference (in distance from the target) between 3D symbology and 2D symbology ($BF_{10} = 2.55$) and between 3D symbology and without symbology ($BF_{10} = 1.71$). At LZ3 there was no reliable difference between levels of symbology ($BF_{10} = 0.23$). These results are depicted graphically as CEPs in Figure 5.2. Further to these results, landings in the 3D Symbology condition were more tightly clustered, exhibiting less variability.

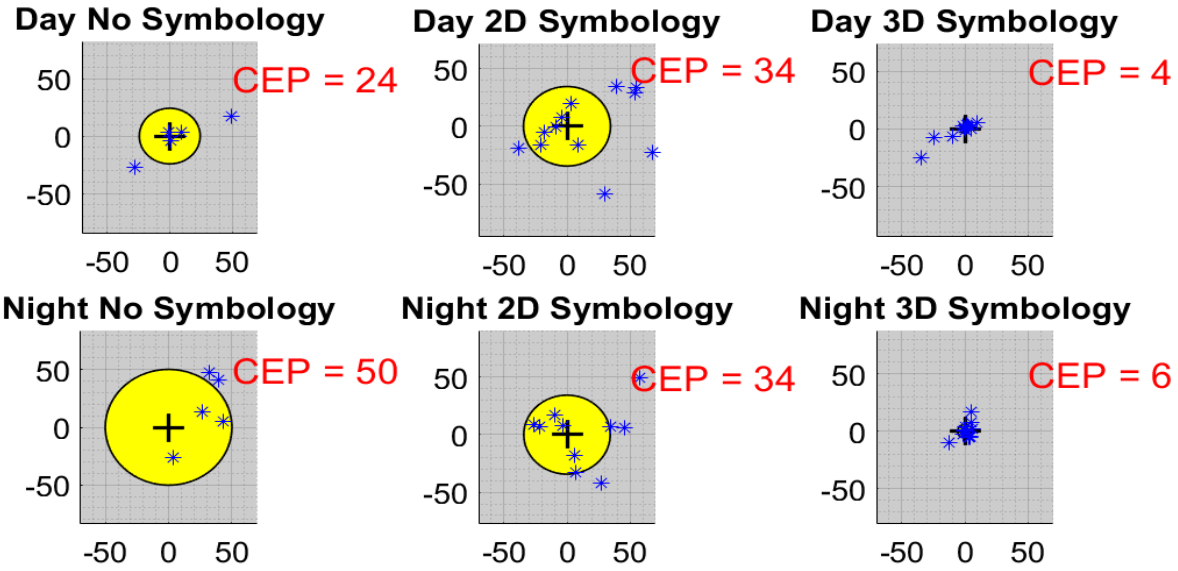


FIGURE 5.2: CEP plots at LZ1 for each environmental condition across all levels of symbology. The cross at the centre of the circle denotes the defined landing zone, with asterisks marking the actual landings in each condition. The yellow circle marks the CEP result for each condition. The CEP value is included in the top right of each plot.

7

The second key performance indicator was comparison to flight targets. The first flight instruction concerned the path between waypoints LZ1 and waypoint E, which followed a river. Pilots were to maintain radar altitude of 200ft and ground-speed of 80 knots. We allowed an absolute deviation of 15 knots, and a +100, -50 ft deviation for altitude (derived in consultation with experienced military pilots). Figure 5.3 shows the proportion of each flight spent outside of these mission-critical parameters (altitude and speed). Bayesian ANOVAs showed a preference for the model which included the effect of symbology, environment and an interaction ($BF_{10} > 1,000$). In both measures (altitude and speed), the 2D symbology condition shows a greater proportion of time outside of the indicated boundaries ($BF_{10} > 1,000$). There is also evidence for an interaction effect between Symbology and environment ($BF_{10} > 1,000$), such that 2D symbology fares much worse in night conditions.

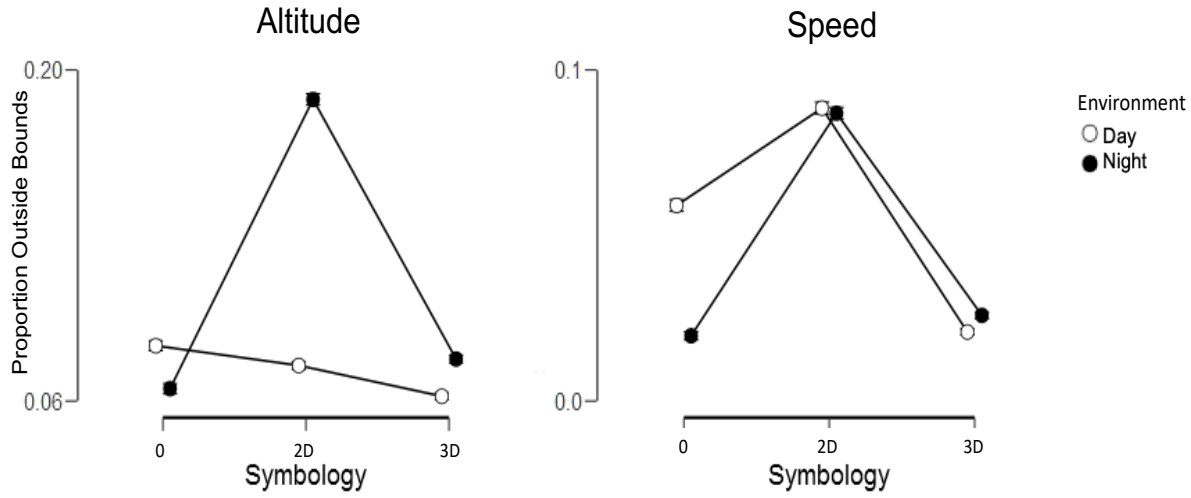


FIGURE 5.3: Left panel; Mean proportion of time that pilots voided the mission bounds for altitude (i.e. flew above 300ft or below 150ft) across participants for the three levels of symbology. Right panel; Mean proportion of time voided the mission bounds in speed (i.e. flew above 95 knots or below 65 knots) across participants for the three levels of symbology. Error bars show the 95% confidence interval, and are too small to see due to the large amount of data – which minimised error.

At LZ2, pilots were instructed to abort landing when they approached very close (a "go around"). For this location, we analyzed minimum altitude and time below the set altitude. The target altitude was 20 feet radar altitude, with "brown-out" occurring when the pilot reached around 120 feet. We conducted Bayesian repeated measures ANOVAs on the minimum altitude reached by pilots for the environmental and symbology conditions, which showed a preference for the model that only included the effect of symbology (BF_{10} = 24.14). The highest minimum altitude was observed in the condition without symbology (M = 33ft), which was higher than the 2D symbology (M = 22ft; BF_{10} = 11.87) and 3D symbology conditions (M = 24ft; BF_{10} = 11.346). No difference was found between the 2D and 3D symbology conditions for this measure (BF_{10} = 0.444). Considering the relative distances of these altitudes from the target altitude, these results show pilots in the conditions with symbology present were able to fly closer to the target than those in the condition without symbology. A Bayesian repeated measures ANOVA on time (in seconds) spent under the target minimum altitude showed a preference for the model which included symbology (BF_{10} = 3.70). Pilots spent more time under the target altitude with 2D symbology (M = 1.29sec) compared to no symbology (M = 0.30sec; BF_{10} = 2.375) and 3D symbology (M = 0.42sec; BF_{10} = 2.613).

Further measures such as flight duration, flight variability across the vertical and horizontal planes were also recorded, but were not reported here, as they fail to add additional insight into flight performance over a longer distance. [8]

From the flight performance data, it is clear that operationalizing optimal flight performance can be challenging. Whilst flight variability provided some insight into performance, and provided data across the entire course of the trial, it is not very informative about flight success and is confounded with highly-trained responses to change flight strategy in different environmental conditions. The CEP plots are limited to only a single value for each flight, yet provide a precise and objective measure of pilot's performance (at least at landing).

5.3.2 DRT

Mean response time was higher in the unsuccessful landing conditions than in the successful landing condition. Bayesian ANOVAs showed a strong preference for the model that included the main effect of landing for log RT ($BF_{10} = 1.588 \times 10^{10}$). Whilst informative in showing the significant increase to cognitive workload during a failed landing, we opted to remove these trials due to the high rate of misses to give a clearer assessment of DRT responses. Pilots were asked to repeat any trial where there was a crash or failed landing.

A two-way Bayesian ANOVA of log RT showed a strong preference for the model that included the effect of symbology (including the condition without symbology), visual condition and the interaction effect ($BF_{10} > 1000$). A comparison of visual conditions showed strong evidence of a difference between the High and Low visual conditions ($BF_{10} > 1000$). A comparison of symbology conditions showed ambiguous evidence of a difference between the 2D and 3D symbology conditions ($BF_{10} < 3$). A two-way repeated measures Bayesian ANOVA of misses was ambiguous, reflecting the relatively small number of missed DRT events (all $BF_{10} < 3$). Figure 5.4 shows log RT for 0D symbology condition in comparison with 2D and 3D symbology across High and Low visibility conditions. The interaction effect

[8] This is a constant challenge in dual-task workload measurement, and as I have shown in earlier chapters, and continue to show in later chapters, is crucial to trade off evaluation. Here it was difficult to find an objective, evidence based (or even agreed upon) criterion for quality flight performance – performance is highly subjective.

is shown here, where the difference between High and Low visibility is exaggerated in 0D symbology, and moves closer together as more symbology is added. This interaction effect suggests that symbology may moderate workload in high difficulty conditions, but may be unnecessary in low difficulty conditions.

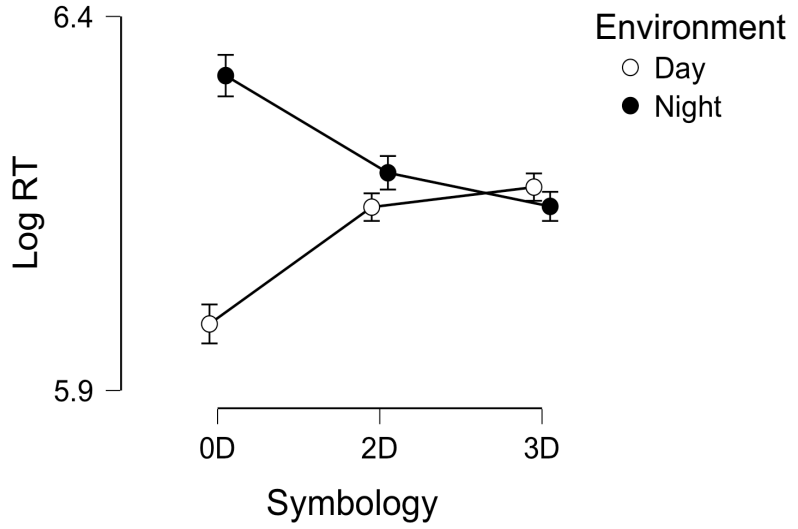


FIGURE 5.4: Mean DRT log RT for environmental conditions across symbology conditions. Log RT is used to normalize across participants. Error bars are 95% confidence intervals.

Additionally, many prior studies have explored both main task performance and cognitive workload, however, there are limited attempts to jointly analyse these. In Figure 5.5 and Figure 5.6, we provide a novel, though rudimentary, combined analysis of both flight performance and cognitive workload. These figures show two conditions (night time with 2D and 3D symbology) for one pilot. We term this analysis a "workload heat map", where a pilot's flight path is plotted in colours that indicate their DRT response latency (calculated as a moving average response time), which is a well established proxy for their cognitive workload. Heatmaps for each pilot in each condition are included in `https://osf.io/2ntxw/`.

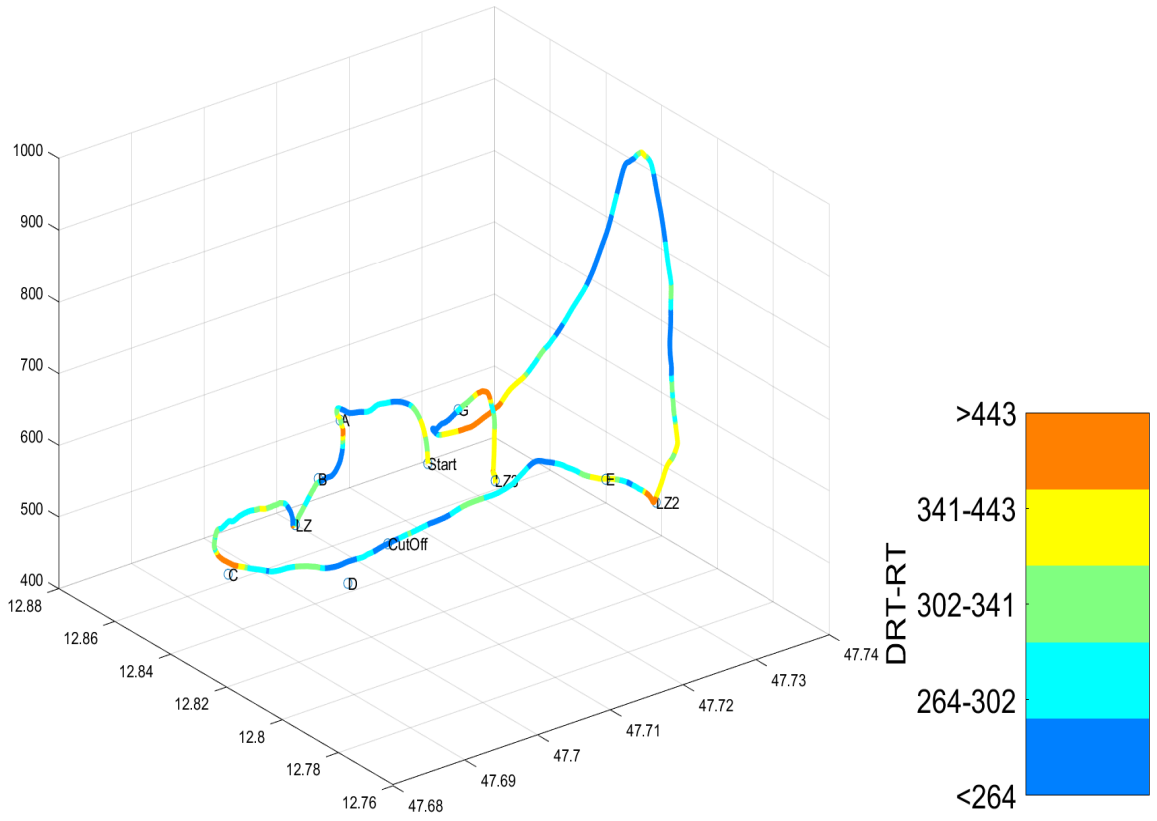


FIGURE 5.5: Workload heat-map for the 2D Night condition for Pilot 2. The x and y axes show latitude and longitude respectively, with the z axis showing altitude. The line displays the flight path that the pilot took. Moving average DRT RT is plotted as colour across the flight for five bins.

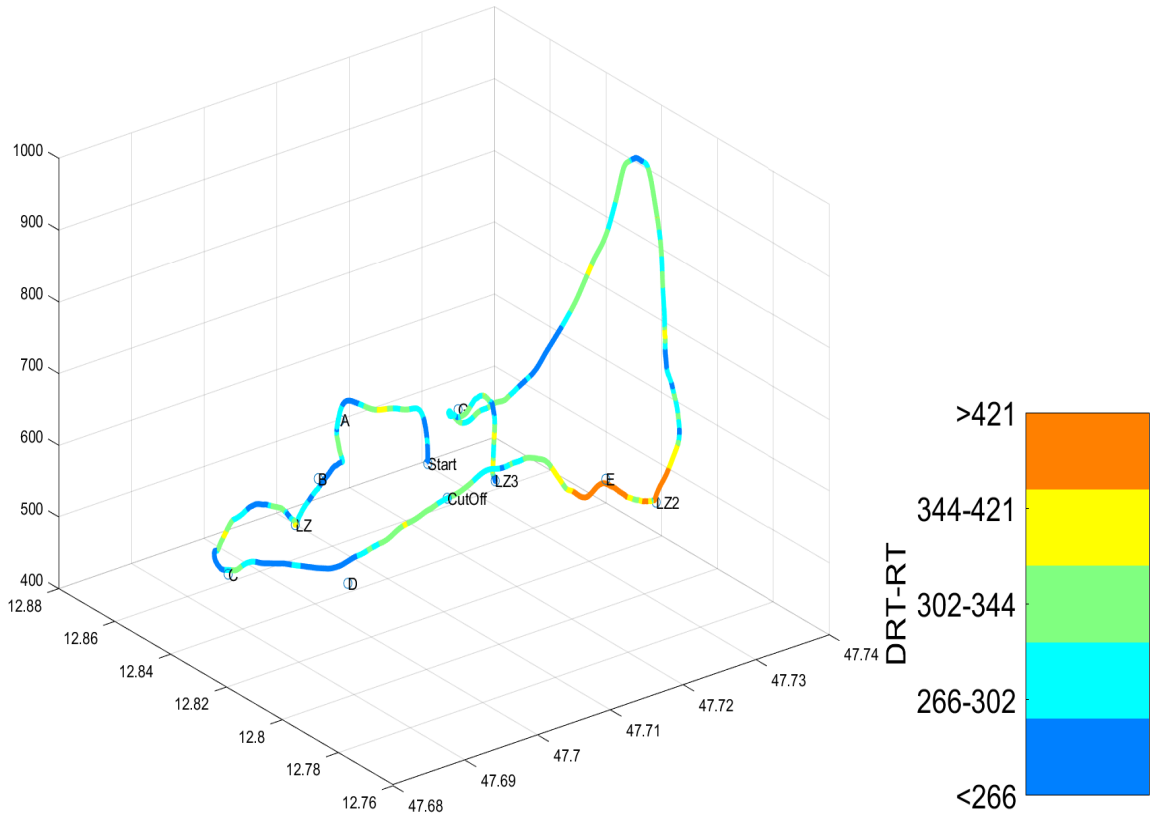


FIGURE 5.6: Workload heat-map for the 3D Night condition for Pilot 2. The x and y axes show latitude and longitude respectively, with the z axis showing altitude. The line displays the flight path that the pilot took. Moving average DRT RT is plotted as colour across the flight for five bins.

5.4 General Discussion

Modern information systems and technological advances aim to assist operators, drivers, and pilots, but often fail to account for the impact on cognitive workload. Complex human machine interactions already present a myriad of information sources, and so before adding to these, it is important to evaluate the impact this information has on the operator (Gerjets et al., 2014; Thorpe et al., 2019). Helicopter environments are highly complex, and so adding more information to the pilots' heads-up-display could potentially prove harmful rather than helpful. The current study used the DRT to evaluate how adding information to a helicopter HUD affected pilots' cognitive workload. Further, we evaluated several flight metrics in an attempt to account for main task performance trade-off.

Results indicated that the DRT was sensitive to workload changes for environmental factors. DRT responses were slower with more difficult flight scenarios, indicating that workload increased in visual conditions of higher difficulty. Contrary to our hypotheses, DRT response times indicated that cognitive workload was relatively unaffected by additional HUD information, with no difference shown.[9] Importantly, an interaction effect was found between symbology and visual conditions, which showed the visual condition having little affect on the 3D condition, but a greater affect on the 2D and 0D condition. This interaction effect is important for future HUD developments, as workload is moderated by symbology in various environmental conditions, which could potentially add unnecessary workload. The interaction effect between symbology and environment provides evidence of a telling story for the importance of user interface evaluation across a variety of conditions. Figure 5.4 shows this interaction which includes the effect of the between-subjects "no symbology" condition. Noticeably, the difference between High and Low visual conditions in the "no symbology" condition provides an insight into the utility of symbology. This provides evidence for the need for an adaptive interface, as, in clear day conditions, adding symbology increases workload - i.e. where it is not necessary. However, in the night conditions, adding symbology actually lowers workload in comparison with "flying blind". [10]

[9] In regards to earlier results, this is similar to those shown in Chapter 3, where there was no difference shown between DRT stimulus type. Here however, in keeping with distraction literature, this finding shows that the more dense information – 3D – had seemingly no distracting effects on workload.

[10] Throughout this thesis, i have identified several results patterns where I highlight that increased workload may or may not be acceptable given context. In this context, it is clear that landing the aircraft is more important than workload – as the workload effects are minimal at most.

Unlike Coleman et al. (2016) and Strayer et al. (2017), our results show that the additional information given to operators may not necessarily cause workload increases. Compared to driving contexts, the helicopter context appears to be much more cognitively demanding, and so this lack of difference in symbology conditions may be a result of a ceiling effect on workload. Alternatively, it may also be plausible to posit that the extremely experienced and highly trained helicopter pilots may be able to more efficiently allocate cognitive resources in order to overcome potential distractors. Research in driving literature suggests this is an unlikely explanation, with Cooper and Strayer (2008) showing no effects of practice on participants ability to overcome distractions. This finding could also be explained by the quality of the heads-up display stimuli, with 3D images more readily perceived than 2D (Dan & Reiner, 2017), meaning that although there was a greater amount of information available, it was moderated by how easily it was perceived – and how useful it might have been to the successful completion of the mission. Regardless of these contributing factors, it is clear that the extra symbology does not add additional cognitive workload to experienced pilots in the helicopter simulator. [11]

Aside from effects of the symbology, the current study does show the usability and sensitivity of the DRT in a previously unexplored environment. The reliability and validity of the DRT has been well documented in driving environments to assess the drivers cognitive workload, however, there are limited applications in other scenarios or environments. With DRT results indicating higher levels of workload for more difficult conditions, we provide evidence that results translate across domains and show the applicability of the measure to helicopter simulator settings.

Figure 5.5 and Figure 5.6 (and further figures in `https://osf.io/2ntxw/`) show a novel approach to jointly evaluate cognitive workload across a flight, an analysis we term "workload heatmap". This analysis shows not only the sensitivity of the DRT, but also provides scope for future analysis to track workload across the duration of a task.[12] The "workload heat map" analysis may not be subjected to simple statistical comparison between

[11] This study differs from prior DRT studies in that, rather than purely evaluate distraction, we also evaluate usefulness of the symbology. Here we show that the symbology is useful to flight performance, and this test is critical for dual-task workload evaluation, as often workload may increase, but sometimes, that increase is beneficial. I will discuss this again later in the conclusions.

[12] In addition to this, joint modelling of tasks could potentially solve this problem of considering simultaneous task performance. For this, a framework of joint modelling dual-task workload measures must first be developed.

conditions, but gives a visual reference to workload distributions across the flight path. We view this type of exploratory analysis as useful for future research (in order to manipulate workload) and for developers of adaptive interfaces.

In regards to flight performance, we analysed a variety of measures to form an objective view of flight quality. Initially we used several "gates" to assess flight performance. This is a commonly used technique in pilots' training and offers a good benchmark for pilots to achieve. The gates provide a marker of performance only for a limited number of locations during the flight, and it is difficult to draw conclusions from the specific gate-locations to the entirety of the task. Secondly, we assessed flight variability. We proposed that an optimal flight would follow a smooth trajectory, on both the horizontal and vertical planes, where sharp movements were indicative of performance lapses. However, over such a long and demanding flight path, which required constant positional adjustments, it was difficult to quantify this measure. Although flight variability gave some indication of the flight path, this was more indicative of the strategy taken (for example at night, it is advised that pilots fly lower) than of flight quality (as a good flight given the conditions may require high variability). Finally, we measured landing precision across the different environmental and symbology conditions.

Landings were a key criteria for the development of the added symbology, as it was used to assist pilots in difficult landing environments. The helicopter simulator was modelled on an Airbus MRH90, an aircraft commonly used in combat zones which require multiple and frequent landings. Consequently, the landing performance across the levels of symbology was a key measure of flight performance, despite only offering a single value for the entirety of the flight. The CEP plots presented in Figure 5.2, provide a clear indication of landing performance across landing zones. This analysis showed landings were more accurate for conditions of 3D symbology compared to conditions of no symbology and 2D symbology. This is a main finding for the current study, as it shows that the increased information provides assistance in difficult landing scenarios (such as at night and in brown–out). Results across environmental conditions show little variance within the 3D symbology condition, but degraded visual conditions have far greater impacts on the condition without symbology and the condition with 2D symbology. Although this metric generalizes performance across the

entire flight to a single instance, it is useful when assessing the impacts of increased visual information.

Overall, results indicated that flying in degraded visual conditions led to higher cognitive workload and had a negative effect on flight performance (as indicated by flight variability and landings). Further, flight performance was unaffected by visual degradation when pilots were provided with 3D symbology. Assessing the flight performance in conjunction with the workload measure allows a more in-depth understanding of the effects of added information: 3D symbology adds no measurable workload, whilst assisting the pilots' flight performance. These results indicate that the 3D symbology may not always be useful for pilots, but is beneficial for night flying at no workload cost. These results show the effectiveness of the DRT as a cognitive workload measure outside of the driving environment and highlights the sensitivity of the DRT as a cognitive workload measurement tool for answering previously inaccessible questions.

The current study was limited by the total number of participants and a restricted stimulus set. Future studies should attempt to quantify what exact information is most useful to pilots, or whether certain symbology elements induce extra workload. Further studies should look to assess the impact of this symbology in more advanced simulators through to real-world helicopter contexts, where the difficulty, and realism, of flying is increased. The impact across conditions of workload should also be assessed to understand whether an adaptive interface is more useful, where the level of information is updated given the difficulty of the current task.

5.4.1 Conclusion

The study offers a unique investigation into pilots' cognitive workload in a high-fidelity flight simulator. The analysis combines various flight metrics with simultaneous assessment of workload via the DRT. The analyses are somewhat limited by the lack of clear optimal main task performance definition and conjoint dual-task analysis. Similar to much cognitive workload literature in driving, it is often difficult to operationalize optimal main task performance, or provide a highly sensitive measure of main task performance. We have attempted to incorporate a variety of meaningful flight analysis alongside the cognitive

workload measures to form a more rounded analysis of this exploratory study. The most telling results generally indicated the expected trends, with little flight path variability between conditions, but a greater effect of added information observed in landing data, where increasing the symbology consistently led to more accurate landings. Flight patterns were shown to vary between environmental conditions regardless of symbology, however landings were highly affected by symbology. Furthermore, the workload measures indicated that the increased symbology added no extra workload, and moderated workload in more degraded visual conditions.

Chapter 6

Application of the DRT as a measure of individual differences

6.1 DRT as a measure of capacity excess: Overview

Following on from Chapters 4 and 5, where I show the usefulness of the DRT as a tool to evaluate displays and additional information, the current chapter has two aims: primarily to use the DRT-MOT as a tool to measure individual differences in cognitive capacity excess, and secondly, to further validate the DRT-MOT design by distinguishing between groups in performance. The DRT has previously been used to evaluate sources of potential distraction, and furthered as a useful tool in evaluating usability of designs, but it is yet to be used to establish individual differences. Evidence for between subject capacity differences exist in systems factorial technology literature (Eidels et al., 2010, 2015; Townsend & Eidels, 2011), and further, cognitive workload literature has shown evidence for large individual differences in dual-task performance (Medeiros-Ward, Watson, & Strayer, 2015). Given the nature of the DRT, where multiple responses are collected per subject, we may be able to distinguish between groups (and by extension between individuals) provided that between subject capacity differences exist. It is theorized that cognitive workload is underpinned by individuals' cognitive capacity, so if these differences do exist, the DRT is positioned as a primary tool to assess capacity. When referring to cognitive capacity, I refer to the excess residual resources, a difference between available and allocated resources, rather than capacity in a systems architecture sense. For this reason, I consider the DRT a measure of cognitive resource excess where individual differences in workload under common task load can infer individual differences in cognitive capacity.

To assist in readability of the introduction, below I summarise the main experimental steps taken to provide context. The primary purpose of this chapter, in using the DRT-MOT as a measurement tool, was motivated by an industry collaboration formed with the ADF group. In the current experiment, I had privileged access to collect data from a cohort of highly trained military personnel, who had been shortlisted for a role based on physical and cognitive criteria. This cohort was recruited from the an elite unit in the Australian Defence Force[1] in a candidate selection program for a highly specialised, and sought after, role. The selection program was for the combat controller position – a position which entails air traffic, and ground traffic, tactical control in war zones. This cohort was expected to be more

[1]Due to privacy concerns from this group, I will refer to this group as the ADF group throughout this thesis.

proficient in this task, based on the selection criteria needed to be selected as a candidate, and more motivated to perform in the DRT-MOT paradigm. Initially, I tested whether results of the difficulty manipulation in the DRT-MOT paradigm held for the RAAF cohort, as shown in previous chapters. Secondly, results of the RAAF group were compared to a control student group – where it was expected that the RAAF would outperform students. Thirdly, in the student comparison group, participants were tested either online or in-lab to observe whether results held across platforms. Finally, a cohort of qualified combat controllers were tested in order to assess the external validity of the task. These participants had been in the combat controller role for a number of years prior to testing. These four steps form the basis of this chapter, with relevant literature and context provided below.

6.2 Experiment 4: Differentiating Groups and Individuals

In many psychological constructs, it is common to observe differences between individuals and groups. These differences can come about through a variety of factors, such as the strategy people choose to complete a task, inherent physiological differences, transference from training, motivation and much more. Many experimental paradigms compare groups in order to understand what factors contribute to performance differences. Valid behavioral paradigms are able to quantitatively distinguish between these groups, where differences are known to exist (Davidson, 2014), in domains such as decision making (Heathcote et al., 2015; Wall et al., 2019), response speed (Rabbitt, 1979) and memory (Grober, Buschke, Crystal, Bang, & Dresner, 1988).

In broader cognitive psychology research, group differences have been shown in a variety of behavioural paradigms between control subjects (generally undergraduate students) and subjects of older age groups (Forstmann et al., 2011), personality groups (Evans, Rae, Bushmakin, Rubin, & Brown, 2017), clinical groups (e.g., depression and schizophrenia; Dillon et al., 2015; Heathcote et al., 2015), and groups differing in blood alcohol levels (van Ravenzwaaij, Dutilh, & Wagenmakers, 2012). These groups all tend to show a cognitive deficit, where in these examples, response time and accuracy is impaired – likely due to

underlying cognitive processes which differ from the the controls. In the current experiment, I had privileged access to a RAAF combat controller selection group of participants who had been selected based on cognitive and physical factors. In testing this group, there are two main differences which exist from previous forms of known-groups testing: first, the RAAF group are assumed to show a cognitive advantage rather than deficit and secondly, this difference is not "known" but rather assumed. In evaluating the DRT-MOT task within this framework, I assume that, if the RAAF group do show a higher performance in comparison to student controls, then this corresponds with an increased ability *and* provides validation of the DRT-MOT to distinguish between groups. In regards to cognitive workload and capacity – the construct being measured by the DRT-MOT – there is precedence to suggest that differences exist between groups and individuals.

Medeiros-Ward et al. (2015) have shown evidence for people who are much more efficient than the average person at multi-tasking, using their "super-tasker" paradigm. The super-tasker paradigm is a short dual n-back task, requiring participants to attend to both auditory and visual stimuli whilst remembering response rules. The researchers showed that a subset of their participants were able to perform the dual-task at the same accuracy as the single n-back task, hypothesizing that these individuals were "super-taskers" and had an innate ability to multitask. This could be due to a higher degree of cognitive control, an inherent ability to more easily switch between tasks, greater experience with multitasking, more experience with mentally manipulating the type of information presented in the n-back task or a higher overall cognitive capacity. Importantly, the paradigm is able to identify the high performers (super-taskers). This literature links closely with cognitive workload, as individuals who exhibit super-tasker profiles are likely to have higher cognitive capacity, or greater efficiency in a workload paradigm.

Group differences in cognitive workload have also been observed in prior behavioral research. Watson and Strayer (2010) highlighted a cluster of outliers on a difficult distracted driving task, whose performance was much higher than the average – they showed no apparent performance decrease in cognitively overloading scenarios. Watson and Strayer (2010) note that the reasons for the heightened ability of super-taskers is difficult to quantify, but is likely an inherent trait or may be trained from completing highly demanding tasks on a

regular basis. This evidence is complimented by research using Systems Factorial Technology where individual differences in cognitive capacity are shown (Townsend & Eidels, 2011; Townsend & Wenger, 2004a) and results from further DRT paradigms, evidencing workload differences between individuals (Conti et al., 2012). Additionally, Jaeggi et al. (2007) showed neural differences between individuals when performing at a high capacity, or exceeding their cognitive capacity, giving evidence that capacity differences may be underpinned by physiological mechanisms. Alternatively, Vidulich and Wickens (1986) highlight that increased performance could be due to motivation factors, where higher motivation leads to increased concentration or performance value.

In addition to differentiating between groups, many cognitive tasks highlight differences between individuals. The study by Medeiros-Ward et al. (2015) is a prime example of this, where the dual n-back paradigm accurately differentiates participants in effective multitasking. In prior experiments, results from the DRT-MOT paradigm has indicated some individual differences – which is shown in the previous chapter (where some individuals performed at a high level despite the impairing text assistance and others performed similarly well with and without assistance). Individual differences are commonly observed across many cognitive behavioural paradigms, however, disentangling ability from noise is often difficult. In the DRT-MOT paradigm, if two individuals are performing the MOT task with the same accuracy, then their DRT results provide an indication of the underlying workload differences between individuals. This trade-off in performance is important in understanding results of the DRT-MOT.

The primary aim of the present chapter is to use the DRT-MOT as a tool to differentiate individuals. This aim was motivated by our collaboration with the ADF group to assist them with combat controller candidate selection. It was proposed that the DRT-MOT task be used in addition to pre-existing selection metrics as a measure of residual cognitive capacity. The specifics of these selection metrics have been withheld here due to privacy concerns, however they mostly entailed several days of intense testing of physical and cognitive capabilities, as well as assessing fatigue management, resilience and emotional intelligence skills. The construct of cognitive capacity is thought to be of high importance to the combat controller role. Thus, it was important to select candidates who showed high performance

in the DRT-MOT, or, more importantly, selected candidates should perform above a certain level.

In order to achieve these outcomes, the experiment included four main testing stages. Initially, the combat controller candidates completed the DRT-MOT task. For this analysis, it was expected that similar trends would be observed for DRT-MOT results as previous experiments (i.e. workload captured as difficulty increased). Secondly, I compared a control group (students) to the RAAF group on the same task. This second step was used as a "known-groups" testing stage, where it was expected that the combat controller candidates would outperform students based on their motivation to perform and their cognitive ability. Furthermore, aside from showing differences between the groups – i.e. the combat controllers are different to students – it was important to validate the design. To do this, I tested two student cohorts; an online cohort and an in-lab cohort. Prior research has indicated that results from cognitive paradigms hold regardless of environment - online or in-lab (Dandurand, Shultz, & Onishi, 2008; Gosling, Vazire, Srivastava, & John, 2004; Krantz & Dalal, 2000; Meyerson & Tryon, 2003). In the current experiment, I expected to see similar performance between online and in-lab participants. Further, if participants *do* differ in performance across testing platforms, the RAAF group can be compared to the in-lab participants to control for this factor (as the RAAF group were tested in-lab). Finally, to test the external validity of the DRT-MOT paradigm, I tested a cohort of qualified combat controllers, who had completed the training and had experience in the role. Comparing the candidate RAAF cohort to the qualified RAAF cohort, provided a form of validation for the task as a *selection metric* – i.e. qualified personnel performed highly. The results of this comparison are useful in selection decisions – i.e. a benchmark.

In the current experiment, I used the DRT-MOT paradigm as shown in Chapter 3 and Innes and Kuhne (2020). Participants completed three levels of difficulty of the MOT. I hypothesized that for all groups, response time and miss proportion in the DRT would increase with MOT difficulty. I also hypothesized that MOT accuracy would decrease as difficulty increased. For the groups, I hypothesized that the RAAF group would outperform the student group in both tasks. I also expected to see the in-lab and online student cohorts performed equally . Finally, I expected the qualified RAAF cohort to show similar results to the candidate RAAF cohort, and outperform both student cohorts.

6.3 Experiment 4 - Method

6.3.1 Participants

In total four cohorts completed the task. The breakdown of these cohorts and respective groupings can be viewed in Table 6.1. The RAAF group was comprised of two cohorts of Royal Australian Air Force personnel; one who were selected as candidates to train for a combat controller course and the other who were qualified combat controllers. There were 53 candidate RAAF participants who completed the experiments in three different sessions. The candidate RAAF cohort were given no incentive to complete the study, however, their results were used in the selection process. The qualified combat controller cohort was comprised of 12 personnel who completed the task as part of their professional development. The online cohort consisted of 63 University of Newcastle psychology undergraduate students who completed the task in their own time online. The Student cohort was also comprised of undergraduate psychology students from the University of Newcastle, however, these students completed the task at the same time at the University, similar to the the RAAF personnel. There were 26 participants in this cohort. Both student cohorts received course credit for completing the study. In total, 154 participants completed the task. Eight subjects (five online, two students, one RAAF) were removed due to computer errors in recording data (where false alarms were wrongly recorded – four participants) or poor performance (DRT miss proportion greater than 50% or MOT accuracy less than 50% for either of the two least difficult conditions).

6.3.2 Tasks

Participants simultaneously undertook the DRT-MOT paradigm. The MOT was displayed on a computer in front of the seated participants, and the DRT was displayed on the screen as a red frame around the MOT display area – identical to the DRT stimulus used in Chapters 3 & 4. There were three levels of workload in the MOT; 0, 1 or 4 dots to track, which was manipulated within-subjects. All participants completed the same task.

6.3.3 Procedure

The RAAF group completed the task on a computer simultaneously in a room of 20-30 computers (as data was collected over several intakes, the room size often varied, however, participants in each intake completed the experiment in the same room). Participants were given individual identification numbers which were used to conceal identity. Participants were given instructions on screen which first introduced the DRT procedure. Participants completed a practice block followed by nine test phase blocks. Each block consisted of ten trials of the MOT, with the exception of the practice block which only consisted of three trials, of random difficulty. Within each test block, all of the trials used the same number of dots to be tracked: either 0, 1 or 4. Each of these levels of load was used for three blocks, giving a total of 30 MOT trials for each load condition. Participants were given breaks between blocks, and the total time taken to complete the experiment was between 1-1.5 hours. The students who completed the study in-lab followed the same procedure as the RAAF group. An experimenter was present to supervise students. The online cohort all received the same instructions and carried out the experiment following the same procedure as the RAAF and student cohorts, but completed the experiment in their own time and with their own computer, without the presence of an experimenter.

Cohort	Group	Number	Location	Reimbursement
Combat Controller Candidates	RAAF	53 (1)	In lab	part of training course
Qualified Combat Controllers	RAAF	12	In lab	professional development
In lab students	Student	26 (2)	In lab	course credit
Online students	Student	63 (5)	Online	course credit

TABLE 6.1: Participants breakdown showing the numbers of participants from each group and exclusions in brackets. Also shown in the table is reimbursement for each cohort.

6.4 Results

The present analysis follows Section 3.2.4.

6.4.1 General Results

The study was treated as a two-way design, with the within-subjects variable of difficulty (0,1,4) and between-subjects factor of group (student or RAAF). The online and student cohorts were grouped as there was no difference observed between these cohorts for all dependent variables (as shown in Section 6.4.4). For the DRT, response time and proportion of missed responses were assessed, whilst for the MOT response time and accuracy were assessed. Mixed, repeated-measures Bayesian ANOVAs were conducted for each of the above measures of interest – between subjects factor of groups and within subjects factor of difficulty. The analysis was conducted using the "BayesFactor" package in R, with default priors (set at 0.707) (Morey & Rouder, 2013). For the ANOVA results, I will again refer to $BF_{Inclusion}$ (from the "bayestestR" package) as the amount of evidence that data are likely under a model containing a given predictor compared to models without this predictor. Results of the Bayesian ANOVAs are presented in Table 6.2.

Overall, both groups performed the MOT task well, with a mean MOT accuracy of 88.07% ($SD = 15.55\%$) and a mean MOT response time of .805 s ($SD = .695$). Figure 6.1 below shows both the change in MOT performance across the levels of difficulty and the difference in performance between the groups. Mean accuracy declined as difficulty increased for both groups, with the RAAF group exhibiting higher accuracy, and mean MOT response time slowed as difficulty increased, with a crossover effect shown between the groups with the effects of difficulty. Bayesian ANOVAs confirmed these trends, as shown in Table 6.2, which indicated strong evidence for the effect of group and difficulty on MOT accuracy, and strong evidence for the effects of difficulty, group and the interaction on MOT response time. These results indicate that increased difficulty lead to lower accuracy and higher response times in the MOT across groups. Results also indicated strong evidence for the effect of groups, with the RAAF group showing higher accuracy than the students. Furthermore, there was evidence for an interaction effect between group and difficulty on MOT response times, with RAAF response times slowing at a greater rate for the effect of difficulty – being faster than students in the 0 dot condition and slower than students in the 4 dot condition. Furthermore, Bayesian t-tests highlighted the reliability of these results, showing differences between groups across all levels of difficulty for MOT accuracy (0 dots – $BF_{10} = 6.53$; 1 dot – $BF_{10} = 83.09$; 4 dots – $BF_{10} = 49.57$). Bayesian t-tests also showed evidence

for a difference between groups on mean MOT response time for the highest difficulty, with ambiguity shown in the lower difficulty conditions (0 dots – $BF_{10} = 1.04$; 1 dot – $BF_{10} = 0.52$; 4 dots – $BF_{10} > 1000$). These results indicate that the MOT difficulty manipulation showed a change in performance, with performance differences shown across groups.

$BF_{Inclusion}$	DRT RT	DRT Miss	MOT RT	MOT acc
difficulty	**>1000**	**>1000**	**>1000**	**>1000**
group	**>1000**	**396.26**	**>1000**	**>1000**
difficulty:group	47/100	41/100	**>1000**	1.28

TABLE 6.2: $BF_{inclusion}$ factors across dependent variables (columns) for each predictor (rows). $BF_{inclusion}$ with sound, or greater, evidence are shown in bold. $BF_{inclusion}$ shown as fractions represent evidence for null effects of the given predictor. $BF_{inclusion}$ greater than three represent evidence for the effects of a given predictor, whilst $BF_{inclusion}$ less than a third represent evidence against the effects of a given predictor.

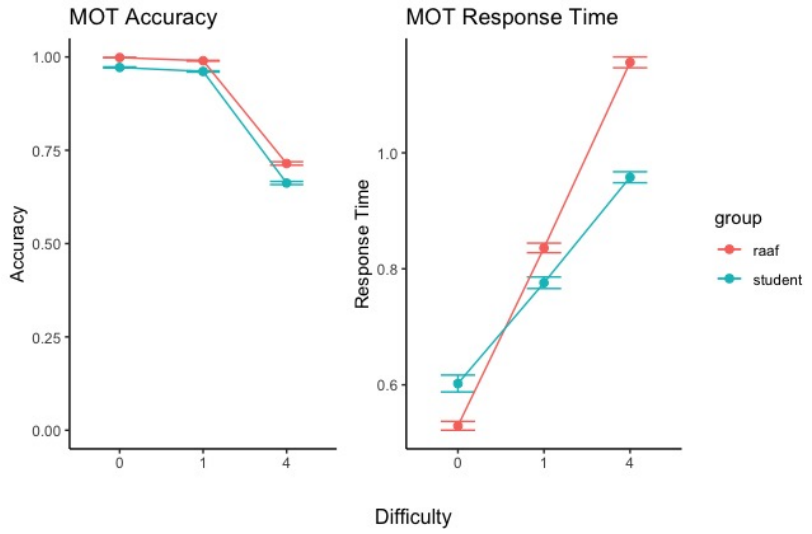


FIGURE 6.1: Performance on the MOT across groups. *Left Panel*: Accuracy in the MOT across levels of difficulty for the two groups. *Right Panel*: Mean response time in the MOT decision phase across levels of difficulty for the two groups. Error bars shown are standard error.

Similarly, the DRT was performed well, with a mean rt of 0.427 s ($SD = .194$) and a miss proportion of 2.8% ($SD = 4.75\%$). Figure 6.2 below shows both the change in DRT performance across the levels of difficulty and the difference in performance between the groups. Figure 6.2 highlights the effects of difficulty on DRT responding, with slower response times and higher miss proportions observed as difficulty increased. Further, Figure 6.2 shows that there was no interaction effect on response time or miss proportion – however, the difference between the groups is observable, with the RAAF group consistently faster

to respond and showing lower miss proportions than the students. A two-way Bayesian ANOVA, shown in Table 6.2 confirmed these trends, with evidence for the the effects of group and difficulty on DRT response time and miss proportion. Bayesian t-tests highlighted the reliability of this result, with a difference shown between groups across all levels of difficulty for DRT response times (0 dots – $BF_{10} > 1000$; 1 dot – $BF_{10} => 1000$; 4 dots – $BF_{10} = 37.47$) and DRT misses (0 dots – $BF_{10} = 40.57$; 1 dot – $BF_{10} = 276.243$; 4 dots – $BF_{10} = 34.83$). Ambiguity was shown for any interaction effects on DRT response times or misses. This indicates that the MOT difficulty manipulation affected DRT results, with the DRT capturing the change in workload across difficulty conditions. Further, these results indicate that the RAAF group had a lower indication of workload and lapses. As opposed to the interaction effect shown above, there was no interaction effects observed in the DRT, indicating that the interaction of difficulty and group on MOT response times may be the result of a strategy difference in responding rather than a result of workload differences.

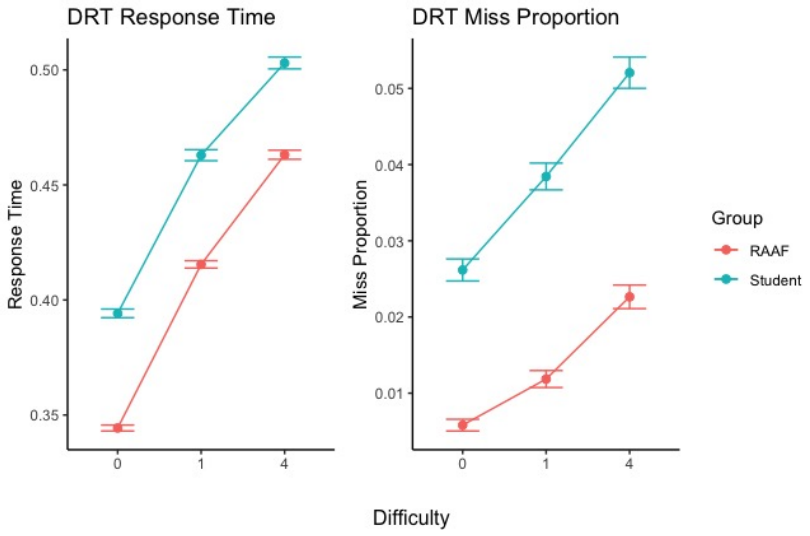


FIGURE 6.2: Performance on the DRT across groups. *Left Panel*: Mean DRT response time across levels of difficulty for the two groups. *Right Panel*: Mean proportion of misses in the DRT across levels of difficulty for the two groups. Error bars shown are standard error.

6.4.2 Individual Analysis

With results indicating that the RAAF group outperformed the student group across all levels of difficulty, it was my next goal to evaluate whether results could distinguish between individuals. The goal of the ADF group was to use the DRT-MOT paradigm

as a further selection metric to add to their battery of tests for the position of combat controller. To assess the cognitive workload capacity, or effective allocation of cognitive resources, of the participants, we need to evaluate results from both the DRT and the MOT. DRT results present an indication of the overall workload of an individual across levels of difficulty. However, this result alone is not sufficient to differentiate between people, as we do not account for performance in the MOT. Performance solely in the MOT, provides no indication of workload. Thus, Figure 6.3 shows performance across the two tasks, for all levels of difficulty, for each individual.

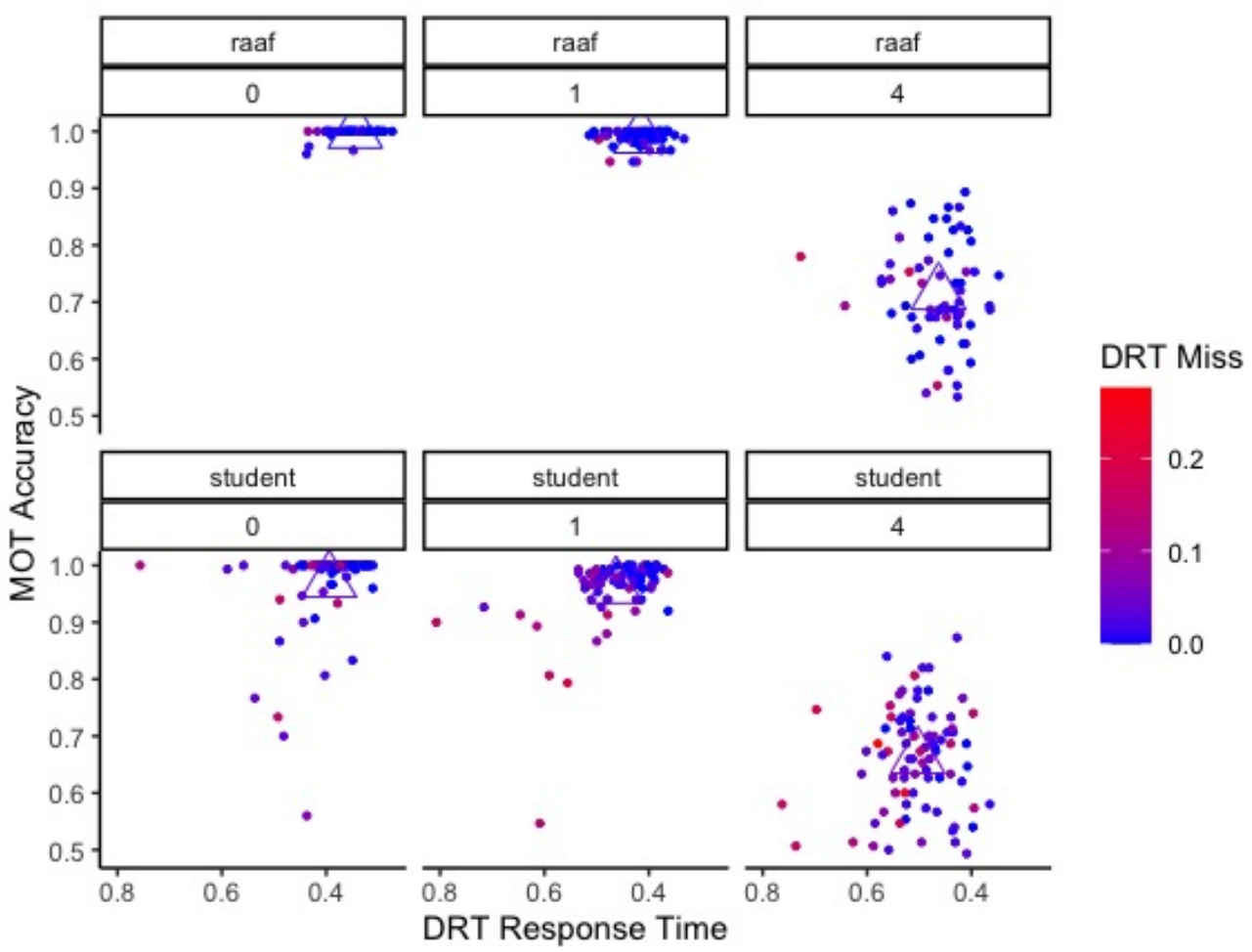


FIGURE 6.3: Individual analysis for the DRT-MOT task across groups and MOT difficulty. Shown on the x-axis is DRT response time (NOTE: Response time increases right to left). The y-axis shows MOT accuracy. Each circular point on the graph represents one individuals data for each group (grouped by rows) and each difficulty condition (grouped by columns). The colour of the dot indicates the miss proportion for that subject. Also included on the graph are the means for each group in each difficulty condition – shown as triangular points. In the graph, high performers are in the top left corner of each panel (i.e. MOT high accuracy, low DRT response time).

For this type of graphical analysis, an assumption is made which fits with the theory of Kahneman (1973). This assumption is that if two participants are completing the MOT task with the same accuracy, then the participant with the lower workload (i.e. slower DRT response times) would have a greater overall capacity. This assumption relies on the limited resources theory, where the main task (MOT) occupies cognitive resources, and the secondary task (DRT) measures residual resources available. With higher accuracy in the MOT and lower DRT response times, participants would exhibit a greater cognitive capacity. This can be seen in Figure 6.3, where participants in the top right of each panel display higher cognitive ability – be it higher overall capacity, better allocation of resources or more effective trade-off between tasks. In the figure, I have also included colouring displaying the amount of misses in the DRT. There are several key trends observable in the data – high performers are seen in the top right of each panel; participants who are valuing the DRT above the MOT are seen in the bottom right of each panel; and participants who value the MOT above the DRT are seen at the top left of each panel. The most apparent condition to separate these trends, and separate people is the 4 dot difficulty.

Evidently, in the RAAF data at the 4 dot difficulty, there is a subset of participants who perform at a very high level in both the DRT and MOT (with MOT accuracy above almost all student participants). Secondly, there is a subset who clearly perform very poorly in the MOT (at around chance level - 60% accuracy). Finally, there are a number of participants who perform at an average level in the MOT, but have slower DRT response times and higher misses – i.e. a higher workload to achieve an average result.

Whilst this graphical analysis is not subject to statistical procedures, it does provide useful information on performance indicative of residual cognitive capacity. This type of analysis may not be useful for distinguishing between high performers, however, it is useful to distinguish high performers from the low performers. With the task indicative of cognitive workload, this is highly useful given the ADF group's needs.

6.4.3 Criterion Validity

Further, to provide a level of criterion and face validation, a qualified combat controller cohort were assessed using the same task. 12 qualified combat controllers, who had

2-14 years of experience in the role, completed the DRT-MOT task following the same procedure as the candidate RAAF cohort.

A series of Bayesian t-tests were conducted to identify differences between the qualified combat controllers and the candidate RAAF cohort. Figures 6.4 and 6.5 show the differences between cohorts. Bayesian t-tests showed evidence for null differences between cohorts in MOT accuracy ($BF_{01} = 5.04$) or MOT response time ($BF_{01} = 4.85$). This was similar for DRT results, with Bayesian t-tests showing evidence for null cohort differences in DRT response time ($BF_{01} = 3.27$) and miss proportion ($BF_{01} = 4.68$). The small sample of qualified combat controllers may have contributed to this result, however, Figures 6.4 and 6.5 show the minimal difference between cohorts.

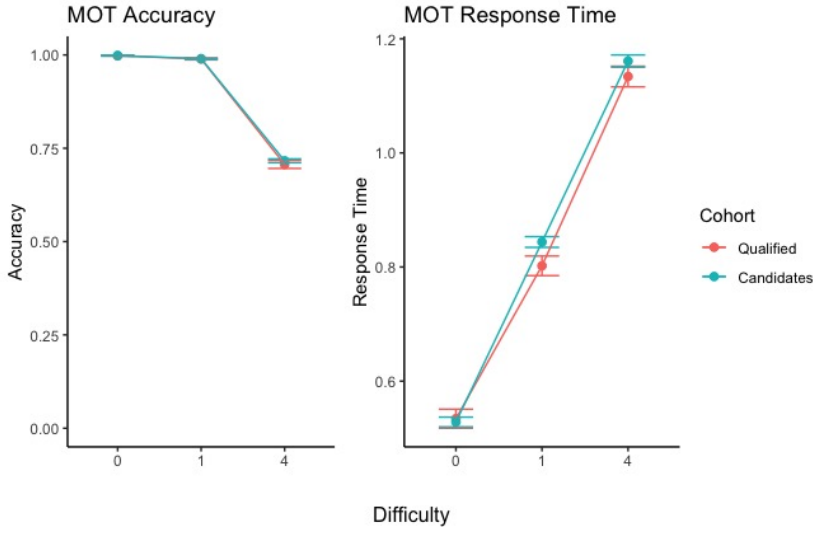


FIGURE 6.4: Performance on the MOT across RAAF cohorts. *Left Panel*: Accuracy in the MOT across levels of difficulty for the two RAAF cohorts. *Right Panel*: Mean response time in the MOT decision phase across levels of difficulty for the two RAAF cohorts. Error bars shown are standard error.

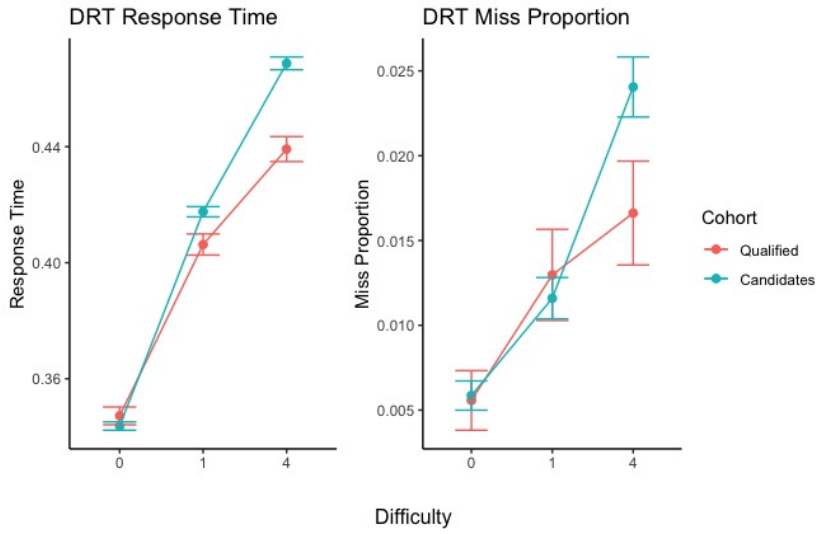


FIGURE 6.5: Performance on the DRT across the two RAAF cohorts. *Left Panel*: Mean DRT response time across levels of difficulty for the two RAAF cohorts. *Right Panel*: Mean proportion of misses in the DRT across levels of difficulty for the two RAAF cohorts. Error bars shown are standard error.

Figure 6.6 shows the individual analysis, similar to that above, however, I include a grey shaded rectangle on each panel. This shaded rectangle occupies the area in which qualified combat controllers scored. This makes comparison between the cohorts more evident, as the outliers in the candidates cohort become more evident. This analysis was similarly useful to the ADF group, as it gave a "benchmark" or reference point to compare candidates to already qualified and tested personnel.

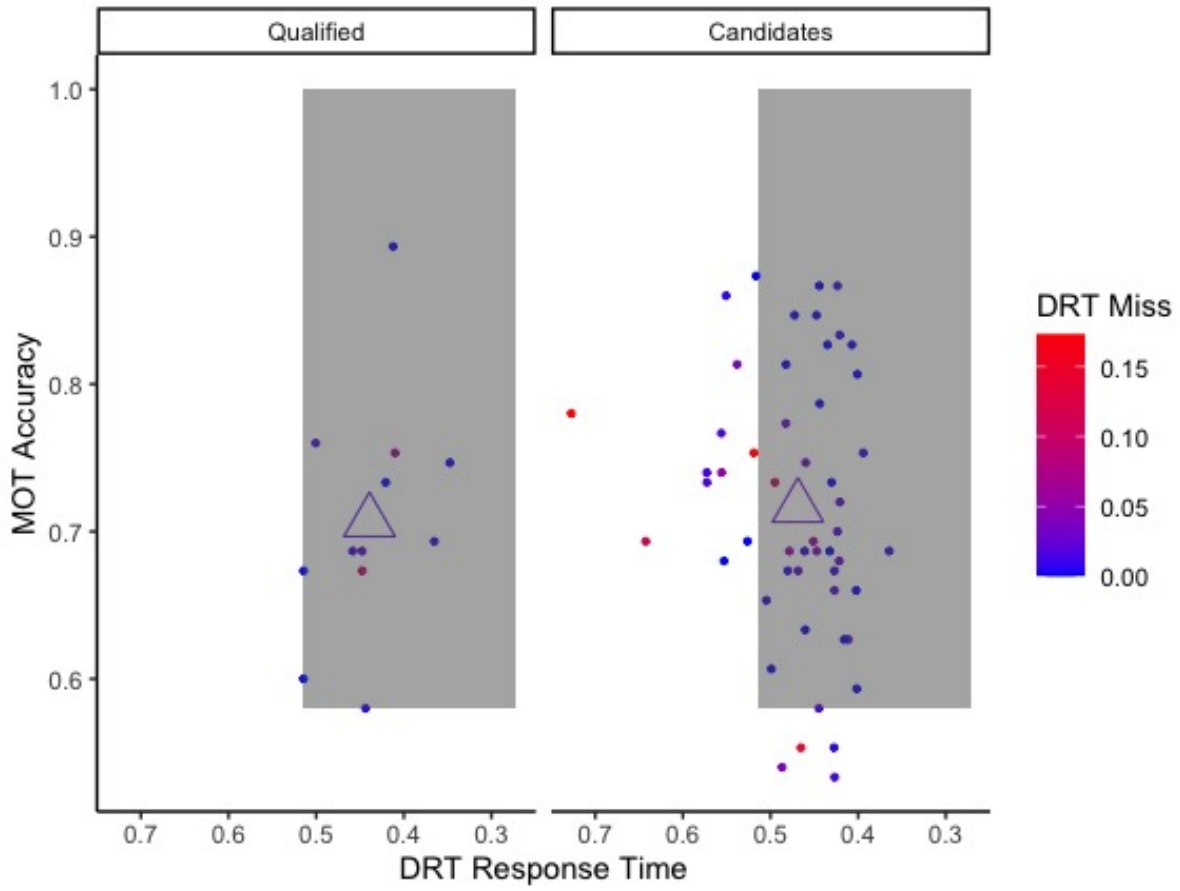


FIGURE 6.6: Individual analysis for the DRT-MOT task across RAAF cohorts for the 4 dots to track MOT difficulty (highest difficulty). Shown on the x-axis is DRT response time (NOTE: Response time increases right to left). The y-axis shows MOT accuracy. Each circular point on the graph represents one individuals data for each RAAF cohort (grouped by rows - left: qualified combat controllers, right: combat controller candidates). The colour of the dot indicates the miss proportion for that subject. Also included on the graph are the means for each cohort in the highest MOT difficulty condition – shown as triangular points. In the graph, high performers are in the top left corner of each panel (i.e. MOT high accuracy, low DRT response time). Also included on the graph is a rectangular annotation, which denotes the area of minimum performance by the qualified combat controllers. This shaded area extends beyond the values exhibited by qualified combat controllers into areas which are of higher performance on both the x and y axes. Participants from the candidate RAAF cohort who fall outside this annotation, were said to fall outside the "benchmark" performance set by the qualified personnel.

Additionally, an analysis was performed on the qualified personnel, using the number of years of experience in the role as an added variable. This analysis showed no effect of years of experience on performance in the DRT-MOT, however, more data is needed to provide conclusive evidence. This data is difficult to access due to the small number of qualified combat controller personnel.

6.4.4 Online vs In lab

Finally, an analysis was undertaken to observe any differences between the online student cohort (online) and the in lab student cohort (in-lab). I initially proposed the in-lab experiment as a more valid comparison level to the RAAF group – who completed the task in-lab. Previous studies have shown similarities across results between online and in-lab environments (Dandurand et al., 2008; Gosling et al., 2004; Krantz & Dalal, 2000; Meyerson & Tryon, 2003). Our analysis extended this finding for the DRT-MOT design. It must be noted in this section that for the removal criteria, a total of five in-lab participants were removed and only 2 online participants were removed.

A series of Bayesian t-tests were carried out to investigate differences between the online and in-lab cohorts for MOT accuracy and response times. Figure 6.7 shows MOT results for accuracy (left panel) and response time (right panel), where there appears to be no difference in performance. Bayesian t-tests confirmed this, showing evidence for a null cohort difference on MOT accuracy ($BF_{01} = 6.19$) and MOT response time ($BF_{01} = 3.86$). This result is in line with earlier chapters and past online vs in lab research (Birnbaum, 2004; Meyerson & Tryon, 2003), but further, shows the external reliability of the DRT-MOT design, with consistent trends shown across environments.

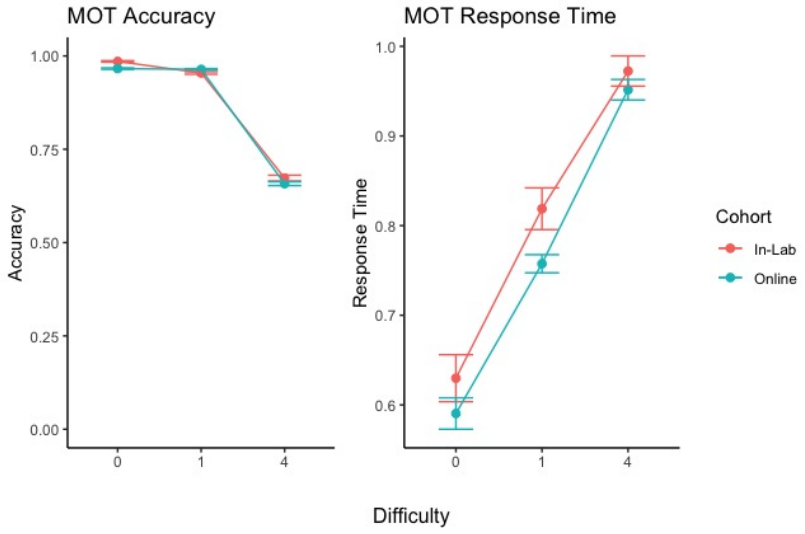


FIGURE 6.7: Performance on the MOT across student cohort environments. *Left Panel*: Accuracy in the MOT across levels of difficulty for the environments. *Right Panel*: Mean response time in the MOT decision phase across levels of difficulty for the two environments. Error bars shown are standard error.

Further, Figure 6.8 shows DRT results for response time (left panel) and miss proportion (right panel), where there appears to be no difference in performance. Bayesian t-tests confirmed these trends, showing evidence for a null difference of cohort on DRT response time ($BF_{01} = 6.63$) and ambiguity for DRT miss proportion ($BF_{10} = 0.50$).

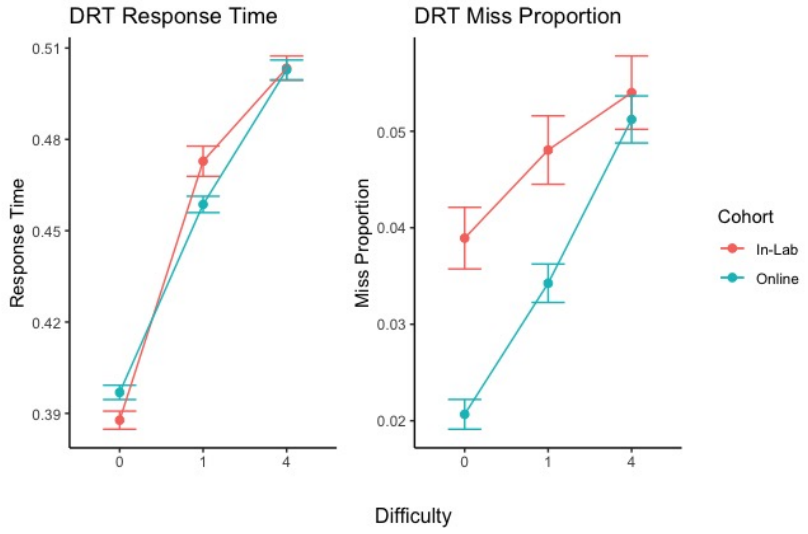


FIGURE 6.8: Performance on the DRT across the two student cohort environments. *Left Panel*: Mean DRT response time across levels of difficulty for the two environments. *Right Panel*: Mean proportion of misses in the DRT across levels of difficulty for the two environments. Error bars shown are standard error.

6.5 Discussion

The current study had several main purposes, the primary of which was to use the DRT-MOT paradigm as a tool – including as a selection metric or to differentiate between individuals. To reach this point, it was necessary to satisfy several other criteria first, as a means of checking the reliability and validity of results. Results generally followed the expected trends, however, a greater range of data (and access to data) is needed before drawing subject specific conclusions. In the following discussion, i will highlight the strengths and weaknesses of the design, however, overall, it is clear that the DRT-MOT design is reliable and valid as a tool to distinguish between groups.

Results across conditions of the MOT followed hypothesized trends, and trends found in previous studies – as difficulty increased (i.e. the number of dots to track), performance in the MOT decreased and cognitive workload increased. The cognitive workload measure – the

DRT – again showed sensitivity to changes in MOT difficulty, highlighting that this measure is reliable, and further, that effects are found across groups. Secondly, group differences were shown in the DRT-MOT design, with the RAAF group outperforming the students reliably across task measures. Furthermore, the online and in-lab student cohorts had comparable performance, showing that the task did not differentiate between groups who should display similar ability.

This group difference test is important in both validating the DRT-MOT design as a tool which can differentiate between groups of different people, and highlighting that expected differences do exist between the RAAF and student groups, but not between the in-lab and online cohorts. It was hypothesized that the RAAF group would outperform students, and this trend was shown across dependent measures, however, it is unclear as to how or why this occurred. Two explanations for this result could be primarily due to motivation factors – as the RAAF group had a greater outcome for good performance (selection into a desirable role); or due to a increased cognitive ability of the RAAF group. Given the potential transference from the daily roles undertaken by the RAAF to the DRT-MOT task, it could be the case that the RAAF group are exposed to more cognitively demanding scenarios more regularly, and this "training" transfers to the task. Explanations for this cognitive ability may be due to an increased cognitive capacity, or a more efficient allocation of resources. These, however, are latent processes which are difficult to observe, and so greater research, including mathematical modelling procedures, should be done to qualify these differences.

The second aim of the study on the road to answering the purpose, was to use results of the DRT-MOT design to distinguish between individuals. Individual differences have long been observed in psychological research, with it generally known that individuals differ in their cognitive ability and underlying neural structures. These differences may be inherent or trained, and may be malleable or fixed. In the current design, the DRT-MOT design was used as a test of cognitive workload ability, where participants were pushed to their cognitive limits. In combining results from both the DRT and the MOT, as seen in Figure 6.3 (which only takes into account performance in the difficult condition), clear patterns emerge. These patterns are primarily related to a task trade off, where as some individuals do poorly in both the DRT and MOT, others only suffer in one domain – essentially trading good

performance in one element of the task to concentrate on the other. In some cases however, participants performance does not suffer in either domain. These "ideal" participants are the type of participants the RAAF were looking to select for their program – participants who showed an increased cognitive workload ability. These participants may have been trading off between tasks at an optimal amount, or had a greater overall capacity, however, they were able to effectively deal with the increased cognitive workload imposed upon them.

It is important to question at this point, does high performance in the DRT-MOT relate to performance in the role in question? To answer this, a further cohort were assessed. This cohort were qualified combat controllers who had been working in the role for at least two years. These individuals were regarded as successful as they had been selected from the candidate pool and had subsequently passed the training and assessment regime. Testing this cohort showed similar results to the candidate RAAF cohort - they outperformed the students across dependent measures and showed similarities to the motivated candidate RAAF cohort. Assessing the combined DRT-MOT results at an individual level was important for external validity and to use as a "reference point" for candidates. Figure 6.6 depicts these results, where it is evident that DRT performance is sacrificed less often than the candidate RAAF cohort, with a form of minimum MOT performance also shown. This cohort of personnel was limited to only 12 subjects, meaning more data is needed before attaining a true benchmark to compare against – however, these results are promising for the external validity of the task as individuals who are employed in the role in question perform at a high level.

In testing the usefulness of the DRT-MOT design as a tool, the preceding results link together to show a tangible story. The DRT-MOT task is able to distinguish between groups, and this performance measure can be used to rule out potentially poor candidates. This study not only shows the usefulness of the DRT-MOT, however, highlights the usefulness of cognitive workload evaluation and the reliability of the DRT as a measure of overall cognitive capacity. This DRT methodology could be applied to a range of tasks used in assessment where roles may require a greater cognitive capacity to be successful. Furthermore, this design may also be used to evaluate clinical patients, where cognitive deficiencies may be resulting symptoms of mental health disorders. Further research is needed in this area, but could provide useful insight to inform treatment and diagnosis.

The current study evidently has several limitations which future work may overcome. Initially, a lack of follow up information on the candidate RAAF cohort makes drawing conclusions on the success of the design difficult. However, participants may fail the selection criteria in another unrelated aspect, which would cloud any conclusions to be made from follow up studies. Secondly, despite candidates undergoing a range of other cognitive testing, access to this data was restricted due to privacy concerns. Access to this data, as well as other demographic information (for both student and RAAF groups), may not only allow insight to correlated cognitive ability, but may also enable a greater picture to be formed of cognitive capacity and what factors inform this ability. Finally, future research should evaluate the test-re-test reliability of the DRT-MOT, as well as observing the effects of cognitive workload training on results. Test-re-test reliability is important when using this paradigm as a selection measure to ensure reliability of the design, while testing candidates pre and post training, would inform our knowledge of factors contributing to increased cognitive ability.

Overall, the current study showed that the DRT-MOT is effective in highlighting differences between the student and RAAF groups, and further, in showing these differences, highlights the validity of the task. Additionally, the combat controller selection panel gained valuable insight from the data when making selection decisions, showing the DRT-MOT as an effective tool to measure cognitive capacity under this task.

Chapter 7

A modelling framework for dual-task cognitive workload measurement using response time distributions.

7.1 Modelling response times

Computational modelling has become increasingly prevalent in cognitive psychology (Farrell & Lewandowsky, 2018). The original goal of cognitive modelling was to offer deeper insight into experimental data in order to infer the previously inaccessible latent processes which underpin behavior. Cognitive modelling has developed into a tool for cognitive scientists, as it allows us to draw conclusions not only about differences between groups, individuals or conditions, but also allows us to make inferences about the processes which may drive these differences.

When applied to behavioural data, there are many examples of successful cognitive modelling approaches which provide greater depth to analysis. Examples include modelling perceptual ability using signal detection theory (Green & Swets, 1966), inferring processes involved in categorization (Farrell, Ratcliff, Cherian, & Segraves, 2006; Nosofsky & Palmeri, 1997), exploring cognitive processes through modelling neural data (Forstmann et al., 2008; Frank, Scheres, & Sherman, 2007), as well as modelling decision making with evidence accumulation models (S. D. Brown & Heathcote, 2008; Ratcliff & Rouder, 2000). Cognitive modelling techniques have enabled researchers to make significant progress in understanding cognition, as we can now make inferences about latent variables such as level of caution, processing speed, sensitivity (to change or detection), and factors that drive preference. Evidently, understanding latent cognitive constructs allows researchers to make better use of data, drawing deeper conclusions and unlocking the so called "black box" that exists between stimulus and behaviour.

The DRT-MOT design used throughout this thesis has several core components which lend themselves to cognitive modelling methods. First, the DRT has simple, quantitative output – response times. Secondly, I have shown accuracy measures from the MOT, however, these fail to provide insight into the tracking *process*, or even the decision process, of the MOT. The tracking process has previously been studied through eye tracking data, however, there is little research into the decision making component of the task. The decision making component however, is a vital component of the task as this stage provides summary measures of the task at hand. Since the decision making component has simple correct/error and response time output, it also lends itself to a modelling framework. The tracking part of

the task is complex and potentially data rich (with technology such as eye tracking) however, the object tracking process requires its own model – many of which have been developed (for examples see; Drew, McCollough, Horowitz, and Vogel (2009); Haritaoglu, Harwood, and Davis (1998); Koller, Weber, and Malik (1994); Rasmussen and Hager (1998)). Whilst informative in some contexts, these models do not permit efficient statistical treatments, and consequently this, and the need for different experimental designs (such as eye tracking and/or closely manipulated object paths), limit the potential relevant applications. Using the data available from the current design, we can infer the "success" of the tracking phase through modelling the decision stage. Here our decision making model can be fit to data, where tracking ability, or success, drives an underlying cognitive processes in the decision phase. Ultimately, through developing a framework which accounts for both tasks, modelling DRT-MOT data can allow us to understand the differences in latent cognitive processes and the relationship between these processes and participant behaviour.

In this chapter, I model DRT-MOT data through a joint modelling framework, using a new sampling process to estimate parameters for two evidence accumulation models. Evidence accumulation models of decision making are ubiquitous in decision making research (Ratcliff, Smith, Brown, & McKoon, 2016). These models propose that for each decision, we perceive decision relevant information (evidence) for decision options which accumulates towards a boundary. Once we have accumulated evidence up to the decision boundary, we then execute this response. This processes has a mathematical theory driven background, however literature has provided evidence for evidence accumulation-like processes within neural data (Forstmann et al., 2010, 2008; Ratcliff, Hasegawa, Hasegawa, Smith, & Segraves, 2007; van Maanen et al., 2011).

For the DRT, I fit a single bounded diffusion model (Heathcote, 2004), where, upon the onset of DRT stimulus elicitation, evidence accumulates towards a single response threshold. Participant response time is determined by the time taken from the stimulus onset to the response threshold with additional time for non-decision processes (i.e. perception, motor execution etc) (Anders, Alario, Van Maanen, et al., 2016; G. Hawkins & Heathcote, 2019; Heathcote, 2004). The rate at which evidence accumulates (i.e. fast or slow) and the setting of the threshold (i.e. more or less cautious) are key components in determining the decision time (Matzke & Wagenmakers, 2009; Ratcliff & Strayer, 2014). For the MOT,

a racing evidence accumulation model is fit, where for each decision in the interrogation phase, an accumulator for each response (indicating whether or not an object was a target) races towards a threshold. Whichever accumulator reaches the threshold first is the "winning" accumulator, and the corresponding response is made(S. Brown & Heathcote, 2005; S. D. Brown & Heathcote, 2008) – for example, in the MOT, there are accumulators for both "yes" and "no" responses, so subsequently, if a "no" response is made, this represents the accumulator which first accumulated evidence to the boundary. Participant response time is made up of decision time (i.e. evidence accumulation to threshold) and non-decision time (S. D. Brown & Heathcote, 2008).

Similar to other evidence accumulation modelling studies such as memory, where strength of the memory trace manifests in drift (Osth, Jansson, Dennis, & Heathcote, 2018; Ratcliff, Thapar, & McKoon, 2011), and preferential decision making, where utility is closely related to drift (G. E. Hawkins et al., 2014), Innes and Kuhne (2020) attempt to map tracking of the target objects on the speed of evidence accumulation (where successful tracking leads to faster processing). Further, Innes and Kuhne (2020) show there is a bias to responding due to unequal response proportions. In the MOT paradigm, there are fewer "yes" responses compared to "no" responses – In the 0 dot condition, 0% of responses should be "yes" responses, in the 1 dot condition, 10% of responses should be "yes" responses and in the 4 dot condition, 40% of responses should be "yes" responses. This bias is accounted for in the threshold parameter, where threshold is lower for more likely responses – for example, in the 1 dot condition, the "no" response is much more likely than a "yes" response (9:1) and so the threshold for "no" responses is adjusted to be lower than the "yes" response. Finally, in jointly estimating parameters from both models, the covariance matrix (i.e. the correlation between the parameters within the model) constrains the parameter estimates. From this, inferences can be made about the interaction between parameters of the two models. This "joint" modelling aspect will be further explained below. By building this framework, I propose a method to assess latent variables which contribute to participant behaviour, and see how these latent variables interact.

The following sections detail the background for methodology, models for both components of the DRT-MOT task and the general sampling procedure before explaining the joint modelling framework.

7.1.1 Accumulator Models of Decision Making

Evidence accumulation models of decision-making are increasingly used as psychometric tools to differentiate between latent cognitive processes between groups, individuals and settings. The Linear Ballistic Accumulator (LBA: S. D. Brown & Heathcote, 2008) and Drift Diffusion Model (DDM: Ratcliff & Rouder, 1998) are two examples of evidence accumulation models of simple decisions. These two prominent models began as theoretical tools to understand the processes which underpinned simple decision making - such as perceptual discrimination tasks. However, in recent years, these models have shown applicability to a number of other research questions – including explaining differences in decision making for clinical populations (e.g. people with schizophrenia (Matzke, Hughes, Badcock, Michie, & Heathcote, 2017), ADHD (Weigard & Huang-Pollock, 2014) and people with depression (Ho et al., 2014)), between groups (e.g. between older and younger adults (Ratcliff, Gomez, & McKoon, 2004; Ratcliff, Thapar, & McKoon, 2007) and linking to IQ scores (Ratcliff, Thapar, & McKoon, 2010)), between environments (e.g. in and out of a fMRI scanner (Forstmann et al., 2008)) and between tasks (e.g. Lerche and Voss (2017) investigated parameter differences of a lexical decision task to a recognition memory task and Hedge, Vivian-Griffiths, Powell, Bompas, and Sumner (2019) compared parameters between Flanker, Stroop and random dot motion tasks). The methodology of these latter examples shows similarities to DRT-MOT experiments.

In addition to the expanding range of tasks modelled under evidence accumulation frameworks, there have been increasing attempts to model multiple tasks together to understand the relationship or correlations between latent processes involved across tasks or domains (Forstmann & Wagenmakers, 2015; Turner, Forstmann, Steyvers, et al., 2019). These "joint models" are motivated by the goal of a more complete model of human cognition, which encompasses the entire task space, rather than focusing on one specific component, hence acknowledging that the tasks being studied involve processes operating together, not separately. Joint models are made more accessible and appealing through broader data, higher computer resources and new computational methods (Turner et al., 2019). Joint modelling often refers to neuroscientific data, where neural data (from fMRI or EEG) is modelled along with behavioural data as a means of utilising both outputs and understanding the links between them – i.e. modelling behavioural data affords insight to underlying

cognitive processes and modelling neural data lends insight into the structures and physiology which may be related to behaviour (Forstmann & Wagenmakers, 2015; Palestro et al., 2018). There now exists a range of linking methods for joint modelling, including directed mapping of neural data to inform parameters (or vice-versa) and covariance approaches (for more information see de Hollander, Forstmann, and Brown (2016) and Turner et al. (2019)). Further, joint modelling of behavioural data from multiple tasks can be used to evaluate correlations between tasks, task trade off and estimate similarity of parameters across tasks using a covariance approach (Wall et al., 2019).

Take for example Wall et al. (2019), who used a joint modelling technique to assess model parameters across tasks and contexts. In their first analysis, the researchers modelled data from Forstmann et al. (2008) collected both in and out of a scanner (for methods see Forstmann et al. (2008) and Wall et al. (2019)). The researchers found that the parameters varied in a similar fashion across contexts, and further, showed parameter correlations where they would be expected (for example, the non-decision time parameter was related between contexts). In the second analysis, Wall et al. (2019) investigated three cognitive tasks – a stop-signal task, a visual search task and a match-to-memory task. The LBA was fit to each task within a joint model framework. From this analysis, correlations between parameters of the LBA were shown across tasks, which enables an understanding of common or divergent processes between tasks – for example, non-decision time correlated across tasks with similar response rules and threshold parameters showed correlations across tasks. This type of across task modelling may assist in understanding common cognitive processes between tasks and individuals.

Innes and Kuhne (2020) present a LBA model application to decision-making data from the MOT task (the data is the same shown in Chapter 6). In this paper, the authors showed evidence for performance differences between a military group and undergraduate control group – where the military group were more accurate, a result which was largely underpinned by the military group setting closer-to-optimal levels of caution. There is scope to improve the analysis by Innes and Kuhne (2020) by incorporating data from the DRT into a joint model framework. This analysis has strengths in that it accounts for the whole task space, meaning a model could highlight trade-off and correlation between tasks. More

importantly however, a model incorporating both aspects of the DRT-MOT paradigm could form a fundamental joint model framework for modelling dual-task workload measures.

7.2 Applications and Methods

In Chapter 6, I outlined a method using results from the DRT-MOT to inform an estimate of cognitive workload "ability", where two groups were compared to identify high-performing individuals. It was evident from the results that this estimate may be affected by individual subject effects – such as the strategy each person used, their level of focus and potential variance in non-decision time processes. In the current chapter, using data from Experiment 4, I conduct a similar analysis to that used by Wall et al. (2019), to jointly estimate parameters of the DRT-MOT task between the RAAF and student groups. For the DRT component of the framework, I fit a shifted-Wald model to the response time distributions, and for MOT decisions, I fit a LBA model to responses (similar to that used by Innes & Kuhne, 2020). I do this through a Particle-Metropolis within Gibbs (PMwG) sampling method, which enables parameters across the two models (i.e. shifted-Wald and LBA) to be estimated together. Through joint estimation of the models, group differences and parameter correlations can be estimated unbiasedly and robustly. This joint model framework is made possible through the PMwG sampler, and is an important primary step towards a more holistic cognitive account of dual-task workload paradigms.

7.2.1 PMwG Model Based Sampling

Wall et al. (2019) extended on a new method of model based sampling to estimate parameters within a joint model framework, as outlined in Gunawan, Hawkins, Tran, Kohn, and Brown (2020). This approach used a PMwG sampler in order to estimate parameters of hierarchical models more efficiently than previous Markov chain Monte Carlo sampling methods. Previously, techniques such as Differential Evolution Markov chain Monte-Carlo (Turner, Forstmann, et al., 2013) have been used to estimate such hierarchical models (i.e. accounting for group-level – parameter – estimates as well as individual – random effect –

estimates). The PMwG sampling method also estimates parameter via Markov chain Monte-Carlo, however, the sampling approach is more efficient and robust, allowing exploration of more complex models.

Similar to Wall et al. (2019), I use the methodology of Gunawan et al. (2020), with sampling conducted using the `pmwg` R package (`https://CRAN.R-project.org/package=pmwg`), and guided by the Wall et al. (2019) extension, where parameters for each model (i.e. per task) are combined to form a single parameter vector α, across tasks. This approach means that a greater number of parameters are estimated, however, with PMwG sampling, this is not problematic. In estimating parameters across tasks, there is an inferred relationship between task parameters, and so data from both tasks is used to inform the the whole parameter space. This relationship is implied by the assumption of a multivariate normal distribution for the random effects, with the relationships between parameters captured by the off-diagonal elements of the covariance matrix (Σ). This structure means that we can observe correlations in parameter estimates within, and between, model components.

The two model components and associated parameters are outlined below.

7.2.2 Modelling Decisions of the MOT

To model decisions in the MOT, I followed the LBA structure as specified by Innes and Kuhne (2020). This model included a mean drift rate within conditions for correct responses (i.e. responding "target" when the interrogated dot was a target had the same drift rate as responding "non-target" when the dot was not a target) and a singular error drift rate. Drift was denoted v_c for correct and v_e for the error drift parameter. I estimated one v_e for two reasons; there was limited data in each condition for error responses and to constrain the model (following recent findings from Evans (2020) regarding model identifiability). There were two thresholds estimated for each condition of MOT difficulty – for responding "non-target" – denoted b_{NT} – or responding "target" – denoted b_T. Each difficulty condition included a mean threshold (b_0, b_1, b_4) and a bias parameter b. The threshold bias parameter decreased threshold for "non-target" responses in lower difficulty conditions as the proportion of correct responses was uneven (i.e. "non-target" was the correct response nine out of ten times in the one dot to track condition). Ultimately, this resulted in four threshold

parameters to be estimated – b (the bias parameter), b_0, b_1, b_4 (the threshold for each difficulty, where adding or subtracting b gave the "target" or "non-target" response thresholds respectively). Three other parameters were also estimated: non-decision time T_0, the uniform distribution of start points A and the correct drift error value s_v. These parameters result in a total of 11 parameters to be estimated (A, b, b_0, b_1, b_4, T_0, v_c^0, v_c^1, v_c^4, v_e, & s_v).

7.2.3 Modelling Responses to the DRT

There have been several recent efforts to fit evidence accumulation models to stimulus-response tasks (such as the DRT) in order to estimate latent variables. Ratcliff and Strayer (2014) modelled the closely-related psychomotor vigilance task (PVT) using an evidence accumulation model of "decision making", where evidence accumulates towards a single boundary (similar to the DDM, where evidence accumulates towards one of two boundaries). Similarly, Tillman, Strayer, Eidels, and Heathcote (2017) fitted a shifted–Wald model to DRT decisions, which followed the same principles. The shifted–Wald model is identical to a one boundary diffusion model, where evidence accumulates to a response triggering threshold and the skewed distribution of these response times is shifted by non-decision time. These models tend to have three key parameters – similar to the LBA – which include a drift rate (or rate of accumulation of evidence), a threshold (or decision boundary indicating the level of caution) and a non-decision time (which relates to the time taken to perceive and encode information as well as the time taken to execute the response). Castro et al. (2019) extended on this model, using a Wald-distributed evidence accumulation model which was augmented by an omission probability parameter. In DRT paradigms omissions may help inform cognitive workload estimates, and so including this data in the model may capture key trends that would otherwise be discarded.

From the above examples, several common phenomena are observed. As task difficulty increases, drift rate tends to slow. Drift rate effects are theorised to be linked with the limitations of human cognition, where occupying more resources causes resource depletion elsewhere (Castro et al., 2019). For example a highly demanding driving task may cause depletion of resources for DRT responding. Another common trend observed is threshold variance, where as caution increases, threshold increases (Rae, Heathcote, Donkin, Averell,

& Brown, 2014; Ratcliff et al., 2016). For example, in a difficult driving scenario, a driver may become more cautious, consequently delaying action as they gather more evidence.

These phenomenon are commonly observed in models of simple decision making. Drift rate effects are often observed when higher task difficulty has an impact on how quickly we are able to process evidence for either choice (Donkin, Brown, Heathcote, & Wagenmakers, 2011; Howard et al., 2020; Ratcliff et al., 2016). Threshold effects are often observed when participants are instructed to respond with less, or more, caution, or are given deadlines (where shorter deadlines force participants to be more speedy) (Forstmann et al., 2008; Karşılar, Simen, Papadakis, & Balcı, 2014; Rae et al., 2014). Similarly, in stimulus-response paradigms, difficulty of the task (for example difficulty in perceiving the stimuli as in Howard et al. (2020)), or difficulty of the concurrent task (such as driving in Castro et al. (2019)), is associated with slower drift rates. These findings regarding drift rate suggest that drift may also be closely associated with cognitive workload. Thorpe et al. (2020) showed similar results, noting a change in drift rate as a result of difficulty, but no threshold effects or effects of differing DRT stimulus modality. Alternatively, cognitive workload may impact on an individuals strategy when responding, as Tillman et al. (2017) found. Tillman et al. (2017) modeled DRT responses in relation to conversations with a passenger while driving. In conditions where individuals were in conversation with passengers, Tillman et al. (2017) showed that threshold increased. This result suggests that increased workload may affect the individuals level of caution, which may vary between individuals.

In the current model application, I model DRT data from Experiment 4 by fitting a shifted–Wald model to participant response times. The shifted–Wald model fitted here also includes a "trigger failure" weighting parameter – similar to Castro et al. (2019). In the DRT, a miss is classified as a failure to respond before the next onset of the DRT signal or any response greater than 2.5 seconds from stimulus onset. In DRT modelling literature, there are two contrasting views on the cause of missed trials: failure to reach threshold or trigger failures. The failure to reach threshold account posits the idea that in trials where participants fail to respond, their accumulation of evidence never reaches the response boundary. In this case, responses could occur at any time in the future, tending towards infinity. Analysis following this theory means misses are removed from the modelling analysis (Howard et al., 2020). The trigger failure account however, posits the idea that on

a proportion of trials, participants fail to *start* accumulating evidence (Castro et al., 2019). The trigger failure parameter is modelled in a mixture distribution framework, where misses are assigned the trigger-failure estimate and provide a weight on the response likelihoods for hits (i.e. $p(x|\theta) \times (1 - pr(TF))$, where $p(x|\theta)$ is the density of the observed data x given the model parameters θ and $pr(TF)$ is the probability of a "trigger failure"). Castro et al. (2019) show evidence against the "failure to reach threshold" account, noting that this would be observed by less right-skewed response time distributions. Further, Castro et al. (2019) show that responses after 2.5 seconds are *less likely* than miss proportions suggest.

Drawing on previous modelling studies employing this practise, I fit a shifted–Wald model with eight estimated parameters. These parameters included the threshold parameter b, drift rate parameter v (one drift for each difficulty), trigger failure weighting parameter f (one for each difficulty) and a non-decision time parameter (i.e. the "shift" for the Wald model) T_0. This gave the vector of eight parameters to be estimated: (b, v_0, v_1, v_4, T_0, f_0, f_1 & f_4).

7.3 Joint Modelling of Experiment 4

A joint modelling approach (as used in Wall et al., 2019) was used for the DRT-MOT paradigm in Experiment 4. Parameters were estimated for the MOT (using the LBA model) and the DRT (using the shifted–Wald model) as outlined above. The full vector of 19 (log-transformed, except for probability parameters f which were probit-transformed) parameters was estimated per participant as a random effect vector, with a multivariate normal prior distribution assumed across participants. The prior for the mean vector of the multivariate normal distribution follows Wall et al. (2019), where I assume this vector is another multivariate normal distribution with a mean of zero and covariance matrix as the identity matrix. The prior for the covariance matrix followed that of Huang and Wand (2013), a mixture of inverse Wishart distributions whose mixture weights were according to an inverse Gaussian distribution. We used settings as specified by Huang and Wand (2013), and by extension Gunawan et al. (2020) and Wall et al. (2019), as these lead to marginally uninformative priors on correlations coefficients. All other PMwG sampling details using the PMwG are according to those in Gunawan et al. (2020).

For PMwG sampling, I used the `pmwg` R package (`https://CRAN.R-project.org/package=pmwg`). For the RAAF group, burn-in was set to 3,000 iterations with 1,000 particles per participant and $\epsilon = 0.0001$, and for the student group, burn-in was set to 100 iterations and 1,000 particles with an $\epsilon = 1$. Adaptation for both fits was set to 100,000 iterations (finishing after around 60,000 iterations for each group) with 2,000 particles per participant and epsilon of 0.0001 and sampling was set to 5,000 iterations with 500 particles and $\epsilon = 0.25$. The value of ϵ shrinks the variance of the proposal distribution in order to increase the probability of a new particle having a high likelihood, thus improving the acceptance rate at the cost of slower coverage of the posterior distribution. It is often used when the parameter space is large (Gunawan et al., 2020; Wall et al., 2019). The burn-in stage differed for the student group in order to have more unique samples in the adaptation stage, which are used to create the efficient proposal distribution in the sampling phase.

The joint model was fit to data from Experiment 4 separately across groups (i.e. the RAAF and student groups were estimated independently using the same model). For each group, participants with MOT accuracy less than 60% (less than chance in 4 dots to track level) in any condition were removed (7 RAAF participants, 21 students). For the outcomes of the joint model estimates, I will show several analyses, all conducted across the two groups (RAAF vs students); descriptive adequacy for MOT decision data and DRT response time data (across groups); parameter estimates for both tasks; correlation matrices. The main focus of this analysis is to estimate both the correlation between parameters of different tasks – to observe the trade-off across cognitive processes for individuals – and to identify between group correlation differences – to observe the differences in trade-off (or ability) across groups. It was expected that the groups would show similar patterns of correlated parameters and further, that parameters of the same constructs (such as drift and non-decision time) would be related across tasks.

Revisiting the parameters of the joint model, I assume a single drift error rate in the LBA as well as single "mean" threshold parameters and single non-decision time parameters for both the LBA and shifted–Wald. The model allows LBA threshold differences and correct drift rate as well as shifted–Wald drift rate and go-failures to vary across conditions. The alpha vector of parameters is represented as:

$$\alpha_j = (A^{lba},\, b^{lba},\, b_0,\, b_1,\, b_4,\, v_c^0,\, v_c^1,\, v_c^4,\, v_e,\, T_0^{lba},\!s_v,\, b^{wald},\, v_0,\, v_1,\, v_4,\, T_0^{wald},\, f_0,\, f_1 \;\&\; f_4)$$

7.3.1 Results

7.3.1.1 Model Descriptive Adequacy

Figures 7.1 and 7.2 show the fit of the joint model for both RAAF and Student data (which were fit independently). Evidently, for both groups, DRT RTs and accuracy are well described by the shifted–Wald model component, while the LBA described MOT decisions well for most conditions. A slight misfit was observed for "yes" responses in the 1 dot condition, where the model underestimates RAAF and student performance, likely due to the low number of responses observed. This misfit is similarly observed in Innes and Kuhne (2020). Further, the model slightly overestimates RT and accuracy in the MOT for "yes" responses in the 4 dot condition. Further plots of model fit can be seen in Appendix B.

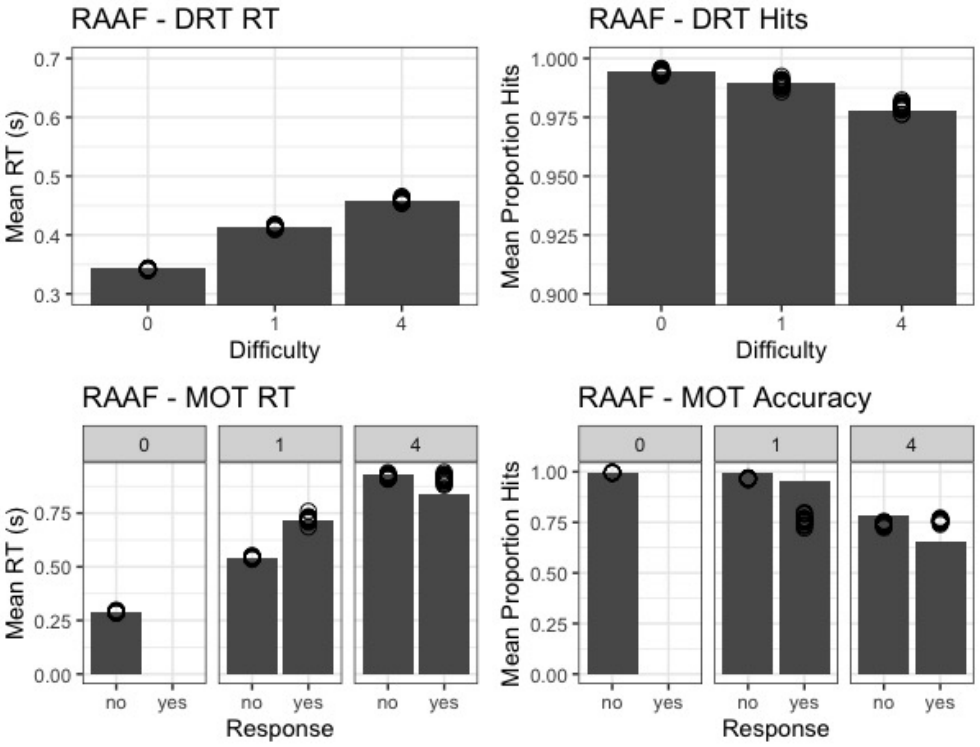


FIGURE 7.1: Descriptive adequacy of the model for RAAF data from Experiment 4. The means from the 20 posterior predictive samples are shown as circles, and the data is plotted as bars. The top row shows DRT data (RT and hits) and the bottom row shows MOT data (RT and accuracy). The top row shows difficulty level across the x-axis, and the bottom row shows difficulty as facets, with response type ("yes" or "no") on the x-axis.

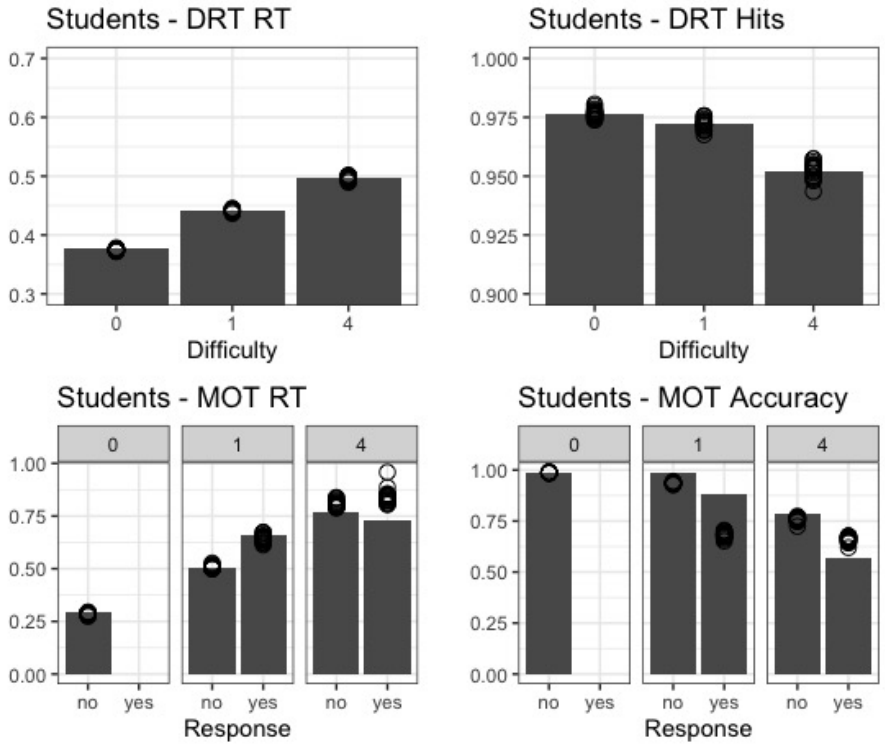


FIGURE 7.2: Descriptive adequacy of the model for student data from Experiment 4. The means from the 20 posterior predictive samples are shown as circles, and the data is plotted as bars. The top row shows DRT data (RT and hits) and the bottom row shows MOT data (RT and accuracy). The top row shows difficulty level across the x-axis, and the bottom row shows difficulty as facets, with response type ("yes" or "no") on the x-axis.

7.3.1.2 Model Results

Table 7.1 shows the posterior estimates for each group across models. Each row shows the mean (and 95% credible interval) of the group level parameters for each task (DRT by shifted–Wald and MOT by LBA), observed from 5,000 posterior samples using the PMwG sampler.

For both groups, shifted–Wald drift and LBA correct drift declines as difficulty increases. This drift effect is expected, with the difficulty level likely affecting the quality of evidence accumulated. As noted in Innes and Kuhne (2020), this effect of difficulty shares parallels with response time modelling in memory research in that the memory trace, or in the MOT the "target trace", drives drift rates. In the memory domain, a stronger memory trace underpins faster drift rates, leading to faster, more accurate responses. In the MOT, a strong target trace, for example when the single target has been easily tracked throughout the trial, would drive faster drift rates for "target" responses, whilst the inverse (i.e. all

distractors in this example), would drive faster "non-target" responses. In more difficult conditions (i.e. more objects to track), this trace weakens as targets are lost or uncertain, leading to slower, less accurate responses.

TABLE 7.1: Mean (and 95% credible intervals) of the estimated posterior distributions for the shifted–Wald and LBA parameters from Experiment 4. Parameter values are shown as "untransformed" for ease of interpretation. Estimates are rounded to two decimal places, except in instances of ambiguity.

		RAAF	**Students**
shifted–Wald			
	b	0.99 (0.90 , 1.09)	0.86 (0.79 , 0.93)
	$v^{(0)}$	5.81 (5.39 , 6.22)	4.44 (4.13 , 4.76)
	$v^{(1)}$	4.06 (3.78 , 4.34)	3.30 (3.10 , 3.51)
	$v^{(4)}$	3.38 (3.14 , 3.68)	2.73 (2.55 , 2.92)
	T_0	0.17 (0.16 , 0.18)	0.17 (0.16 , 0.19)
	$f^{(0)}$	0.0006 (0.0002 , 0.0013)	0.005 (0.002 , 0.009)
	$f^{(1)}$	0.002 (0.001 , 0.005)	0.015 (0.010 , 0.023)
	$f^{(4)}$	0.008 (0.004 , 0.013)	0.028 (0.018 , 0.040)
LBA			
	A	0.21 (0.15 , 0.28)	0.17 (0.13 , 0.24)
	b	0.97 (0.90 , 1.05)	0.75 (0.66 , 0.84)
	$b^{(0)}$	1.29 (1.10 , 1.49)	0.96 (0.87 , 1.05)
	$b^{(1)}$	0.75 (0.71 , 0.78)	0.74 (0.71 , 0.78)
	$b^{(4)}$	0.54 (0.51 , 0.57)	0.60 (0.57 , 0.64)
	T_0	0.05 (0.04 , 0.07)	0.03 (0.02 , 0.04)
	$v_c^{(0)}$	2.32 (1.97 , 2.69)	2.33 (2.03 , 2.67)
	$v_c^{(1)}$	2.19 (2.06 , 2.32)	1.78 (1.66 , 1.91)
	$v_c^{(4)}$	1.40 (1.30 , 1.50)	1.21 (1.12 , 1.31)
	v_e	0.09 (0.07 , 0.13)	0.07 (0.06 , 0.10)
	sv	0.74 (0.70 , 0.78)	0.68 (0.63 , 0.73)

Similarly, the f parameter estimates for the DRT increase as difficulty increases, showing that omissions become more likely in harder tracking conditions. This is a similar finding to work by Castro et al. (2019), and shares parallels with the drift effect. Trigger failures are responses where the accumulator fails to start, and henceforth an omission occurs. In the DRT-MOT, the increase in DRT trigger failures with increasing difficulty could be the result of higher workload, where a lack of attentional resources leads to instances where participants fail to detect the stimuli or trigger response mechanisms, as they are focused on tracking.

Finally, the b bias parameter reliably declines as difficulty increases, which is in line with expectations given the probability of alternate responses for each condition. Recall that the b bias parameter is either added to or subtracted from the threshold parameter for the "target" or "non-target" accumulators respectively to account for the probability of responses. For example in the 1 dot condition, "non-target" responses account for 90% of the data, whereas in the 4 dot condition, "non-target" responses account for 60% of the data, hence a bias is present, making the "no" response more likely in lower difficulty conditions. Mean parameter estimates reflect this intuition, with both groups showing a decrease in the level of bias as response proportions tend towards evenness.

Between groups, there are several notable findings (which can be seen in Appendix B Figure B.4 for group differences in the shifted-Wald and Figure B.5 for group differences in the LBA). To assess differences between groups, I compared 95% credible intervals for each parameter – as shown in parentheses in Table 7.1. In the DRT component, the RAAF personnel set more cautious thresholds, but have much higher drifts – in fact drift rates for the RAAF in the hardest (4 dot) condition are similar to students' drift rates in the less difficult (1 dot) condition. The RAAF group also have far fewer trigger failures. Parameter estimates from the shifted–Wald model component highlight that overall, the RAAF are simply better than students at the DRT – RAAF participants able to show greater caution to the signal, but also a greater processing speed of information. However, the LBA component must also be accounted for to understand the between task dynamics across groups.

In the LBA component of the model, group differences are less evident. Drift rates and non-decision times are comparable between the groups. The main difference student and RAAF groups is observed in the threshold parameter estimates. RAAF participants appear to set higher overall thresholds, as indicated by the b parameter. The magnitude of difference between the thresholds of competing accumulators, indicated by the b-bias parameters, appears to have an interaction effect with participant groups, where the RAAF group set higher thresholds in the 0 dot condition, but lower thresholds in the 4 dot condition. This threshold magnitude result implies more optimal responding (due to uneven response proportions) by the RAAF group, who show a greater threshold change than students across difficulty levels – that is, in easy conditions, RAAF participants are less cautious for more probable ("non-target") responses and more cautious for less probable decisions ("target"),

and in hard conditions, the RAAF group are similarly cautious for the similarly probable responses. The student group do not appear to be as sensitive to this response probability change and consequently adjust their thresholds less across difficulty conditions. Further, the RAAF group show higher correct drift rates in the 1 and 4 dot conditions, which, when coupled with their more optimal threshold setting, indicates the key mechanisms underpinning their superior performance: set optimal caution levels *and* have higher processing ability.

Overall, it is clear that the RAAF group outperform the student group, and these mean parameter estimates provide an indication as to what processes underlie this behaviour. In addition to these findings, the "joint" component of the dual-task paradigm is important in understanding parameters that are related or traded off between tasks. Table 7.2 shows the lower rectangles of the covariance matrices for each group – i.e. the parameter correlations *between tasks* (Appendix (B) tables show the full covariance matrices for each group). Evidently, there are only several parameters which reliably correlate between the tasks, however, these can inform the model outcomes.

Table 7.2 shows that in the LBA (rows), correct drift correlates positively with drift estimates from the shifted–Wald (columns). Further, correlations are also observed between LBA correct drift and shifted–Wald trigger failures, showing that when drift in the MOT is high, participants exhibit lower misses in the DRT. Appendix B furthers the validity of this analysis, with strong covariances shown within component parameters; i.e. LBA parameters correlate with LBA parameters.

Between groups, the trigger failure parameter appeared to have stronger negative correlations with LBA error drift for the student group than the RAAF group. This negative correlation means that error drift rates increased with less trigger failures and vice-versa, in what could be some form of task trade-off as expected (i.e. if a participant is concentrating on the DRT, they would have less trigger failures, but may have more MOT errors). This could indicate further differences between the groups, where the RAAF participants were able to trade-off between tasks without losing key information. The correlation in shifted-Wald non-decision time and LBA correct drift (particularly in the 4 dot condition) for the RAAF group, may also be indicative of differences between the groups in levels of engagement. The correlation is lower in easier conditions, however, this could be the result of more accurate drift estimation in this condition, compared to the 0 and 1 dot conditions, which have limited

responses for some design cells. This is predominantly observed in the RAAF group, which could suggest the RAAF group show more similar performance.

TABLE 7.2: Mean estimates for the lower rectangle (shifted–Wald with LBA parameters) of the correlation matrix from Experiment 4. Reliable estimates are shown in bold - meaning that generally the distribution of correlation values for these parameters did not cross zero. The full correlation matrices can be seen in Appendix B.

RAAF

	B	$v^{(0)}$	$v^{(1)}$	$v^{(4)}$	T_0	$f^{(0)}$	$f^{(1)}$	$f^{(4)}$
A	0.07	0.01	-0.07	-0.13	-0.02	0.04	0.08	0.11
b	0.12	0.12	0.08	0.02	-0.08	0.08	0.11	0.11
$b^{(0)}$	-0.08	-0.03	-0.02	0.02	0.03	0.12	-0.06	-0.11
$b^{(1)}$	-0.04	-0.13	-0.15	-0.12	0.02	0.09	0.11	0.06
$b^{(4)}$	-0.02	0.05	0.02	-0.01	0.15	0.02	0.04	0.08
T_0	-0.11	0.14	0.08	0.03	0.14	-0.13	-0.12	-0.07
$v_c^{(0)}$	0.05	-0.05	-0.05	-0.11	-0.04	-0.08	0.13	0.14
$v_c^{(1)}$	0.11	0.08	0.13	0.09	-0.15	-0.03	-0.02	-0.09
$v_c^{(4)}$	**0.29**	**0.29**	**0.33**	**0.27**	-0.25	-0.17	-0.22	-0.17
v_e	0.07	0.10	0.11	0.09	-0.07	0.06	-0.01	-0.02
sv	-0.05	-0.03	-0.04	-0.12	0.03	0.07	0.13	0.08

Student

	B	$v^{(0)}$	$v^{(1)}$	$v^{(4)}$	T_0	$f^{(0)}$	$f^{(1)}$	$f^{(4)}$
A	0.15	0.15	0.03	0.07	-0.07	-0.21	-0.06	0.03
b	0.12	0.15	0.02	-0.11	0.04	-0.12	-0.04	-0.07
$b^{(0)}$	-0.03	0.06	0.06	-0.05	0.18	-0.06	-0.07	-0.13
$b^{(1)}$	-0.09	0.05	0.04	0.06	0.07	-0.10	0.03	-0.04
$b^{(4)}$	-0.22	-0.13	-0.20	-0.05	0.06	0.17	0.22	0.18
T_0	-0.05	0.09	0.16	0.04	0.11	0.15	-0.04	-0.15
$v_c^{(0)}$	0.14	0.18	0.03	0.07	-0.19	-0.14	-0.04	0.04
$v_c^{(1)}$	0.23	**0.30**	0.23	0.08	-0.04	-0.19	-0.17	-0.11
$v_c^{(4)}$	0.16	0.20	0.22	0.09	-0.03	-0.17	-0.12	-0.13
v_e	0.21	**0.33**	0.30	0.28	-0.07	**-0.35**	-0.26	-0.26
sv	-0.04	0.13	0.03	-0.02	0.12	-0.17	-0.17	-0.11

7.4 Discussion

Using the methods developed by Gunawan et al. (2020) and Wall et al. (2019), I apply a joint modelling framework to data from the DRT-MOT task. This is a novel analysis approach for MOT data and, with the integration of the DRT task, shows a comprehensive insight into the latent cognitive processes which underpin task performance and workload

allocation. In jointly estimating parameters for respective models, I provide a method to jointly model dual-task workload measures. This enables us to investigate parameter changes related to MOT difficulty levels for both tasks, as well as the trade-off between tasks and the correlation between model parameters. Thus, this analysis goes beyond the depth of previous chapters, in that I am able to simultaneously account for performance across both tasks and the latent mechanisms underpinning behavioural outcomes.

In regards to workload, and the between task trade-off, there is evidence to suggest that some individuals are generally better at the task, optimising decisions and resource allocation. One would expect that if attention was paid to the DRT, MOT performance would decline – likely observed in a decreasing LBA correct drift parameter. At the same time, we would expect that DRT performance would increase, with trigger failures decreasing and shifted-Wald drift increasing. This is an intuitive example scenario of a between task trade-off. Surprisingly, as results show in Table 7.2, this is not necessarily the case, but rather the opposite appears to occur. Drift is positively correlated between model components, whilst trigger failure and LBA drift are *negatively* correlated. This result implies that subjects with higher drift in one task, generally show similar high performance in the other task. Further, as performance increases in MOT, trigger failures decrease in the DRT. Correlations are also observed between drift and non-decision times, indicating that the same underlying processes may be responsible for these factors across tasks. This factor may relate to participant engagement, where non-decision times are minimized as participants are more focused and attentive.

As shown in previous chapters, results indicated differences between RAAF and student cohorts. Interestingly, from the model, these differences were primarily observed in the DRT task, where the RAAF group reliably showed higher drift *and* thresholds compared to students. In the MOT, the RAAF group appeared to have a more optimal strategy, changing their threshold bias more substantially for responses in lower difficulty conditions, whilst setting more even thresholds in more difficult conditions. This may benefit performance – where response times are faster and more accurate in easy conditions and slower, but more accurate, in harder conditions – a pattern that can be observed in Figure 6.1.

Results from the joint model indicate a positive start for this type of analysis and show advantages over traditional dual-task workload measure analysis. Future extensions

could address several limitations that were encountered in the present analysis and data set. MOT response accuracy appears to follow slightly incorrect patterns for the RAAF participants – which could be the result of limited data, as each condition had only 30 MOT trials, and further, response proportions were much lower for easier conditions (for example in the 1 dot condition where the misfit is greatest, there is only a 10% chance on each trial of giving a "yes" response). The MOT does appear somewhat challenging to model – particularly as it is difficult to account for the actual tracking behaviour (i.e. why a participant is performing poorly could depend on a range of circumstances). These difficulties could be overcome in future by including more difficult conditions in the MOT, including a contaminant parameter in the model, or through using another, more simple cognitive task to manipulate workload (such as that in Thorpe et al. (2020)). Further, the covariance matrices showed few significant correlations, meaning that generally the posterior distribution of correlation values crossed zero, and so whilst the means suggest correlations, these may not be statistically reliable. This result may suggest that variance within each group was large, with minimal data across this spread, or alternately, may suggest that there is a limited relationship between the tasks. Future simulation studies and studies with more data in relevant conditions could address this issue, whilst future joint modelling research should attempt to incorporate groups in the same fit.

It is also important to note however, that this model makes only predictions, drawn from the data, about the underlying cognitive processes. Further, the current model may be overly complex for this kind of analysis. The main factors that I am interested in here could be reduced to drift rate and drift rate variability, meaning the LBA framework, whilst it is able to capture all elements of the decision-making process, is overly complex. In future research, two shifted-Wald models could be used (one for MOT and one for the DRT) to answer more specific, workload relevant questions in regards to this type of design. Future researchers should strongly consider their choice of models to ensure that the model appropriately addresses their main question without over-extending. Finally, it should be added that here, the model is not the process, but merely a description of the possible processes, and so results should be interpreted with more caution than the data provide – especially in regards to cognitive workload measurement.

In estimating dependent parameters between tasks, a more holistic model framework

is explored accounting for the variance and correlations between tasks. Evidently the model shown here provides a sound summary of the data, with parameter estimates highlighting causes for cohort performance differences in both the DRT and MOT. Whist the covariance analysis is somewhat limited by the data, the joint model methodology outlined here provides a novel approach to dual-task cognitive workload measurement.

Chapter 8

General Conclusions

Throughout this thesis, I have examined cognitive workload theory and measurement, an important human factor in many in-lab and real world contexts. Overall, I found that the detection response task (DRT) was a reliable measure of cognitive workload and this held across tasks and designs. Secondly, using the multiple object tracking task (MOT) as a manipulator of cognitive workload was effective across experiments, where increasing the number of objects to track led to a performance decrease in tracking accuracy and subsequent increased workload. Implementing this DRT-MOT paradigm enabled me to explore behavioural outcomes resulting from increased workload across a range of applications. Results showed consistency across experiments; increased workload lead to lower MOT performance and slower DRT response times. With the difficulty manipulation affecting performance on both tasks, we can be confident in our assessment of behavioural outcomes of cognitive workload factors. Further, results from this thesis highlight the reliability of the DRT-MOT paradigm and the scope of cognitive workload applications.

This thesis makes several key contributions to both cognitive psychology literature and human factors literature. In the theoretical stream, I identified a flexible, widely applicable cognitive workload measure, and proposed potential applications to evaluate workload across a range of contexts. It is evident that cognitive workload plays a role in performance in many different fields, however, the accessibility of workload evaluation is often limited and consequently overlooked. I initially discussed the range of cognitive workload measures available to researchers in Chapter 2, and in Chapter 3 I provided validation for an in-lab cognitive workload paradigm. Chapter 3 also highlighted that this DRT extension is able to be distributed online without compromising reliability of data. In Chapters 4, 5 and 6, I implemented DRT methodology in novel environments and for novel purposes. This allows greater insight into potentially critical human factors, and allows researchers to ask new questions. In this methods-focused stream, I provided a framework for assessing cognitive workload in new fields, such as aviation environments as shown in Chapter 5, and for new purposes, such as evaluation of types of assistance as shown in Chapter 4 and personnel evaluation as shown in Chapter 6. Further, in Chapter 7, I proposed a joint-modelling framework for dual-task cognitive workload measurement, a novel modelling application in cognitive workload literature which enables behavioural results to be further extrapolated. The developments in the methodology stream have the potential to broaden the uptake of

dual-task cognitive workload measures in a range of new contexts and for a range of purposes, whilst the analytical methods discussed have potential to extend the scope of cognitive workload research.

Chapters 1 and 2 identified the importance of cognitive workload measurement, with distracted driving research highlighting the importance of this construct. Following the methods of Innes, Evans, et al. (2020), I showed the validity and reliability of the DRT-MOT. Using the MOT as the main task allows participants to be "response-free", but still requires task engagement in tracking the objects. Further, the cognitive workload effects of the MOT are easily manipulated, by increasing the number of objects to track. In Chapter 3, I showed several tests of reliability and validity, which provide evidence that the DRT is consistent across settings, and has sound external and construct validity. This is important in extending the DRT to use in less cognitively demanding laboratory-based settings (in comparison to highly demanding driving settings) and highlights the sensitivity of the DRT to workload change when conditions are carefully controlled and manipulated. As noted in Innes, Evans, et al. (2020), the DRT-MOT paradigm has potential applications, where environmental factors could be manipulated to assess the effects of this manipulation on task performance. However, in this thesis, I used this paradigm to highlight the scope and applicability of the DRT and secondly, used the DRT to evaluate novel research questions - as shown in the methodology section.

In the methodology stream, I extended tests of the reliability and usability of the DRT to new purposes, for example in Chapter 4, assessing the effectiveness of added assistance on MOT tracking accuracy. Previous uses of the DRT have tended to focus on the effects of distraction on cognitive workload as a result of environmental factors or alternate task engagement (Stojmenova & Sodnik, 2018; Strayer et al., 2013; Young et al., 2013). In Chapter 4, I instead evaluated two forms of assistance which *appear* to be helpful for the task, to see how these assistance types affect workload (and task performance). Results from this study identified key patterns in data where the added assistance was either; helpful but costly to workload or unhelpful but not costly to workload. I expect that using the DRT for similar evaluative purposes can highlight similar patterns across environments – where useful information is often costly to attentional resources or alternatively, not costly, but also not useful. Ultimately, this falls on designers and researchers to identify how much these factors

can be balanced against one another. The balance between these factors can be conceptualised in Figure 8.1, where the green cell indicates an ideal cost–benefit trade–off between workload and performance, the red cell indicates a poor example of this trade–off, and yellow cells indicate instances in which context and environmental factors must be accounted for. For example, in some scenarios, adding any workload to the operator may be deemed too dangerous, whereas in other contexts, operator workload may be less crucial and performance could be improved. Each problem poses unique implications which need to be considered within their own context, and so whilst Chapter 4, does not provide a unified solution to this problem, it does provide a methodology to assess such factors. Results from Chapter 4 show a pattern consistent with the red and yellow cells (top right and bottom right quadrants) of Figure 8.1, for the text assistance and reappearing target assistance respectively.

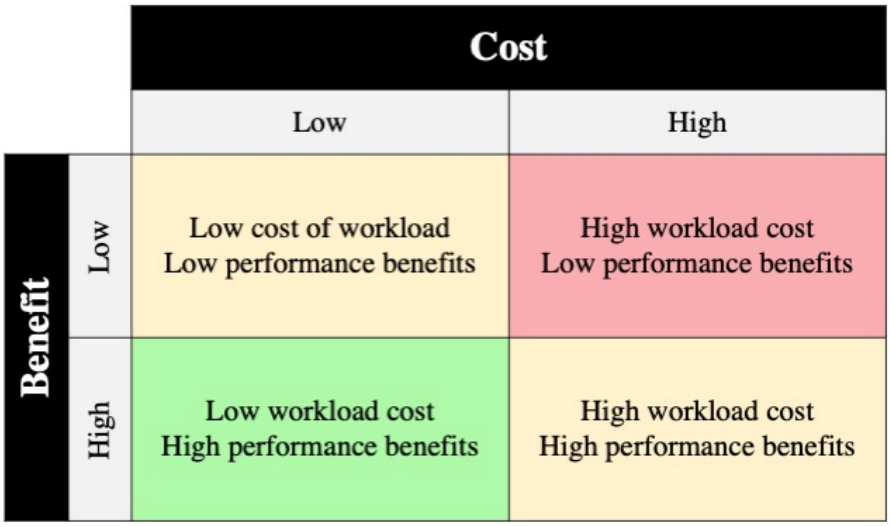


FIGURE 8.1: Table of theoretical trade-off for cost to user workload and benefit to user task performance of added information. Green cells are optimal situations, red cells indicate detrimental situations and yellow cells indicate scenarios where "optimal" is context dependent, for example if workload is less important, then a high workload cost for a high performance benefit is an optimal scenario.

In Chapter 5, I applied such methodology in a helicopter simulator setting to evaluate the effects of heads-up display information. Results from this study showed that increased information (3D symbology) significantly increased flight performance, and came at no significant cost to cognitive workload, as per the bottom left (green) quadrant of Figure 8.1. In this example, whilst the assistance does add workload, this is very minimal, and in this context, performance (i.e. safe landing) has more value.

In addition to these findings, Chapter 6 shows a use of the DRT to inform estimates of individual cognitive workload ability, which could be used for personnel selection. This cognitive workload ability is task specific, where in the DRT-MOT paradigm, if two participants have similar performance in the MOT task, then we can infer that the participant with better DRT performance (lower mean response time) would imply greater "cognitive capacity". Evidently, this assumption is subject to many other factors, such as individual differences in non-decision time and level of caution adopted, however, it provides a useful benchmark to compare potential candidates. In Chapter 6, I had privileged access highly trained RAAF operators – who were used as a "benchmark" – and trainee RAAF operators who were undergoing a selection process. Further, I tested two student cohorts (online and in-lab), who under-performed on the task in comparison to the RAAF participants. This provided a form of "known-groups" testing *and* showed group differences between the students and the RAAF personnel for both DRT-MOT components. In a further test of task reliability, results did not differ between testing environments (online and in-lab) for the student group.

In an extension of the analysis from Chapter 6, I used an emerging sampling technique to jointly estimate model parameters from the task. The analysis in Chapter 7 used the PMwG sampling method (Gunawan et al., 2020) to estimate parameters of the Shifted-Wald model (for DRT data) and LBA model (for data from decisions in the MOT) in a joint model framework, so that covariance could also be estimated. This allowed for a deeper insight into latent cognitive processes and overcomes the limitation of individual strategy and non-decision time differences outlined previously. Here, I found that the military group outperformed the student group in both components of the DRT-MOT task and was able to relate these performance differences to more optimal threshold setting and faster evidence accumulation processes. Further, correlations between tasks showed evidence against a trade off between tasks impacting performance, but rather a facilitation, where good performers in one task often performed well in the other. This could reflect the level of engagement or motivation experienced by participants, and could also encompass their cognitive ability. This analysis showed correlations across tasks and differences between the groups, but most importantly, provided a novel approach to dual-task workload measurement analysis which accounted for both components of task.

Although this thesis offers exciting developments in cognitive workload measurement, it has several limitations. Primarily, this thesis is limited in that data is primarily derived from the DRT-MOT paradigm, and I urge future researchers to investigate other dual-task cognitive workload paradigms. These paradigms could include a variety of standard cognitive tasks as a means of evaluating the workload factors present; or could extend to further practical applications. For practical applications, as shown in Chapter 5, it is critical to have a measure of main task performance in addition to DRT data. Measuring main task performance is important to distinguish between cells of Table 8.1, as knowing workload outcomes alone is often not informative. Furthermore, under the joint modelling framework established in Chapter 7, data from such real–world designs could be extrapolated to provide greater insight into multitasking behaviour. Another limitation of this thesis is that I predominantly rely on DRT data to estimate cognitive workload. In addition to the aforementioned future directions, researchers may look to combine other physiological or neural measures of workload with DRT data to increase understanding of the mechanisms underpinning behaviour. The joint modelling framework outlined in Chapter 7 could also be applied with other workload measures, as outlined by Forstmann and Wagenmakers (2015) and Turner, Sederberg, Brown, and Steyvers (2013), where data from a neural model, for example, informs the behavioural model (or vice-versa). Further, I have recently tested the DRT as a cognitive workload measure in tandem with physiological measures of workload (such as heart rate, blood pressure and arteriole pressure). This experiment is part of a parallel project looking at framing of difficult scenarios, where some individuals experience a state of threat compared to others who experience a state of challenge. Initial results show promising links between measures, however, it is far beyond the scope of this thesis.

In summary, this thesis provides a comprehensive framework for dual-task cognitive workload measurement, and assists in extending dual-task workload measurement purposes, analyses and applications. Throughout the thesis I explore the applicability and usability of stimulus response tasks (specifically the DRT) through a single framework, which is used for a variety of purposes. This highlights the flexibility and reliability of these tasks, and provides a platform for researchers to build on where dual-task workload measurement can be seamlessly incorporated into future paradigms to understand workload effects. I report here on several experimental methods, dual-task analyses and show powerful modelling techniques – the first of their kind in dual cognitive workload literature. From this research, I have made

a substantial contribution to the field of cognitive workload measurement, demonstrated novel applications of these methods and tools and furthered the scope of cognitive workload measurement.

The Candidate extends his sincere thanks to the examiner and appreciates the time they have taken to read this thesis.

Appendices

Appendix A

Chapter 5 Appendix

A.1 Glossary

- **Brown-out** - An instance where dust from below the helicopter is disturbed and rises to an altitude of about 120ft, thereby hampering the view for the pilot.
- **Collective lever** - controls the angle of the main rotor blades, allowing the helicopter to accelerate or decelerate.
- **Cyclic shaft** - changes the main rotors direction in order to change the direction of the helicopters movement.
- **FLIR** - Forward looking infrared radar. A sensor system that uses infrared light to see at night.
- **Ground Speed** - the speed (in knots) that the aircraft is travelling in reference to the ground
- **HUD** - Heads-up display. The information presented in the HUD is overlaid over the environment so that they do not have to shift gaze to perceive the stimulus.
- **Landing zone** (LZ) - a designated point on the map where pilots were to land. The landing zone was clearly marked in the symbology, on the map and by objects in the environment (i.e. the centre of a football field).
- **Radalt** - Radar altimeter measures altitude above the terrain that is currently beneath the aircraft.
- **LIDAR** - Light detection and ranging. A sensor system that uses pulses of laser light to measure variable distance to the ground.
- **Roll** - The degree of sideways movement in the aircraft
- **Pitch** - The degree of forward and back movement of the aircraft
- **Symbology** - The information given to pilots within their heads up display. Includes general flight metrics and more advanced environmental information.

A.2 Full Flight Path

Pilots were seated in the simulator and fitted with the visor and DRT's tactor patch. They were given instructions for responding to the DRT, and for completing the flight task. Three experimenters were present to collect data, with one experimenter collecting DRT data, another updating the parameters of the simulator, and a supervisor. An additional pilot was also present, navigating the participant through the flight as required. Pilots were instructed in the symbology presented in the 3D-symbology condition, and were given time to acclimate with the system. Before the flight commenced, the pilot was given a practice block of DRT trials to familiarise themselves with the stimulus and response button.

In the Day condition, visibility was set at 12,000m, time of day was set at 16:00, FLIR and dust were off. In the Night condition, FLIR was on and was set at 20:00 with FLIR visibility at 2,400m. General visibility in this condition was set at 12,000m, time of day was set at 20:00 and dust was off. In the Low Visibility, Dust condition, the dust appeared at 100m from the ground. Visibility in this condition was set at 1,200m, time of day was set at 16:00, dust was on and FLIR was off.

The flight task was divided into six sections. Way points were placed throughout the map to indicate the key points. Way points were marked on the control panel map and indicated in the symbology (for both 2D and 3D conditions). Section 1 required the pilots to take off from a designated helipad and fly to two waypoints, designated Way-point A and Way-point B. In Section 2, pilots landed at their first LZ, designated LZ 1, which was a flat sandbank. Pilots encountered brownout during this landing. Brownout began at 100ft, with a simulated brownout fully engulfing the virtual aircraft to restrict view by roughly 60ft. Section 3 was a second flight section, in which pilots followed a river through a valley to Way-points C and D, marked on two bridges along the valley, and Way-point E, marked on a church at the end of the valley. Pilots were given ideal speed and height levels of 80kn and 200ft, and instructed to fly as close to these levels as possible during this section. Section 4 required pilots to descend to a LZ, designated LZ 2, which was marked on a triangular brown field. Pilots were instructed to "go around" or abort the landing at height of 20ft. Going below this set altitude in a real-world scenario would be potentially dangerous and could compromise mission objectives. As with LZ 1, pilots encountered brownout, which

was removed when pilots cleared power lines located behind LZ 2. Section 5 was the final flight section, in which pilots ascended and descended a mountain, flying towards Way-point G nearby the take-off helipad. Section 6 was the final landing on the flight deck of a Nimitz-class aircraft carrier. The LZ, designated LZ 3, was the junction of the centre lines of the carrier's straight runway and angled runway. The full flight took approximately 13 minutes to complete. Pilots were seated in the simulator and fitted with the visor and DRT's tactile patch. They were given instructions for responding to the DRT, and for completing the flight task. Three experimenters were present to collect data, with one experimenter collecting DRT data, another updating the parameters of the simulator, and a supervisor. An additional pilot was also present, navigating the participant through the flight as required. Pilots were instructed in the symbology presented in the 3D-symbology condition, and were given time to acclimate with the system. Before the flight commenced, the pilot was given a practice block of DRT trials to familiarize themselves with the stimulus and response button. Pilots were instructed to begin the flight upon responding to the first DRT stimulus they perceived. After completing the first two sections, including landing at LZ 1, pilots were instructed to take off and continue the flight after several seconds on the ground (following standard flight procedures). They then completed the last four sections of the flight.

A.3 Symbology Conditions

- No symbology: In this condition, the pilot was equipped with the HUD headpiece (as shown in Figure A.1, however, it was turned off so that pilots could still see the full disply with no extra visual information.

FIGURE A.1: An example of the simulator setup. The pilot has the head piece attached which displays the HUD information over the simulated environment. In front of the pilot are the electronic map and a multi-function display, which indicated altitude, ground speed, collective power and helicopter roll.

- 2D: In the 2D condition, pilots were equipped with the HUD headpiece which displayed several metrics in their visual field. These metrics included radial altitude, ground speed, heading, distance & direction to the landing zone. An example screenshot can be seen in Figure A.2.

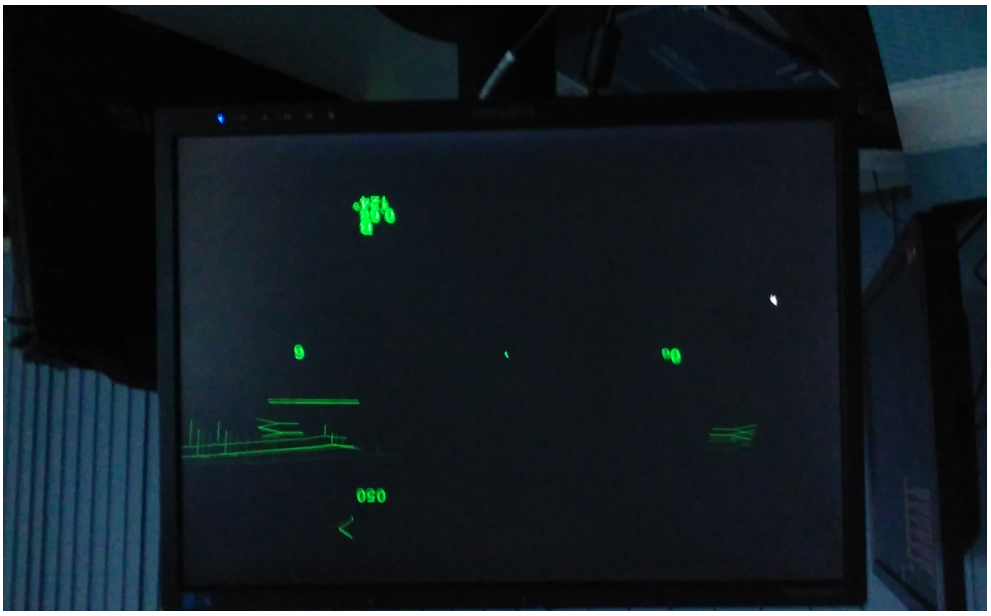


FIGURE A.2: An example of the projections for the 2D symbology condition. The information shown on screen was projected to the HUD in the headpiece worn by the pilot.

- 3D: In the 3D condition, pilots were equipped with the HUD headpiece which displayed several metrics in their visual field, as well as overlaying 3D visual information to the simulated environment. These metrics included radial altitude, ground speed, heading, distance & direction to the landing zone. The 3D information also given to pilots included 3D mapping of landing zones (as seen in Zimmermann et al. (2019)), flight path direction, and visual indication of obstacles (such as buildings and power lines; as shown in Figure A.3).

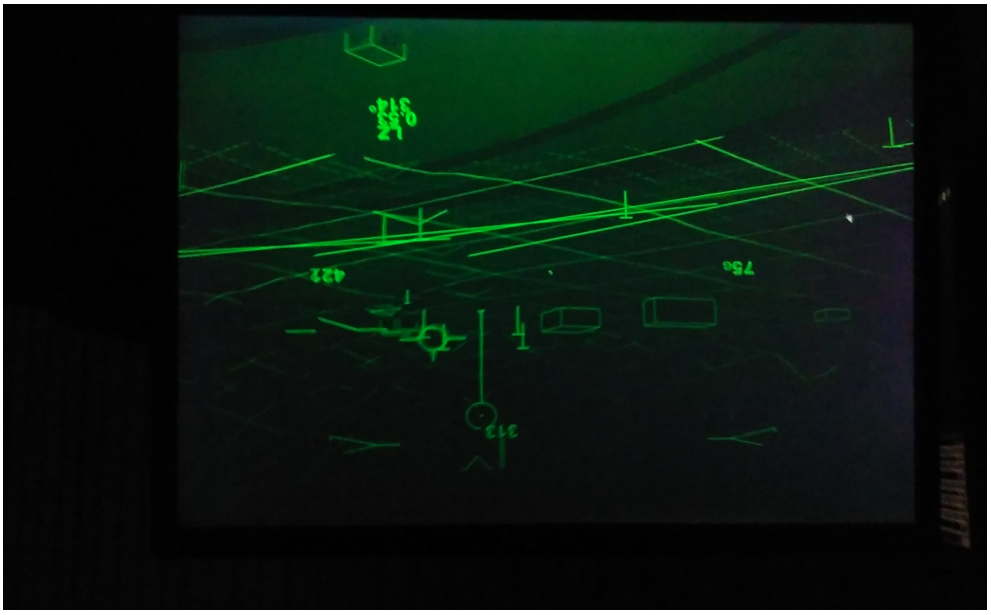


FIGURE A.3: An example of the projections for the 3D symbology condition. The information shown on screen was projected to the HUD in the headpiece worn by the pilot.

Appendix B

Chapter 7 Appendix

B.1 Further Plots of Descriptive Adequacy

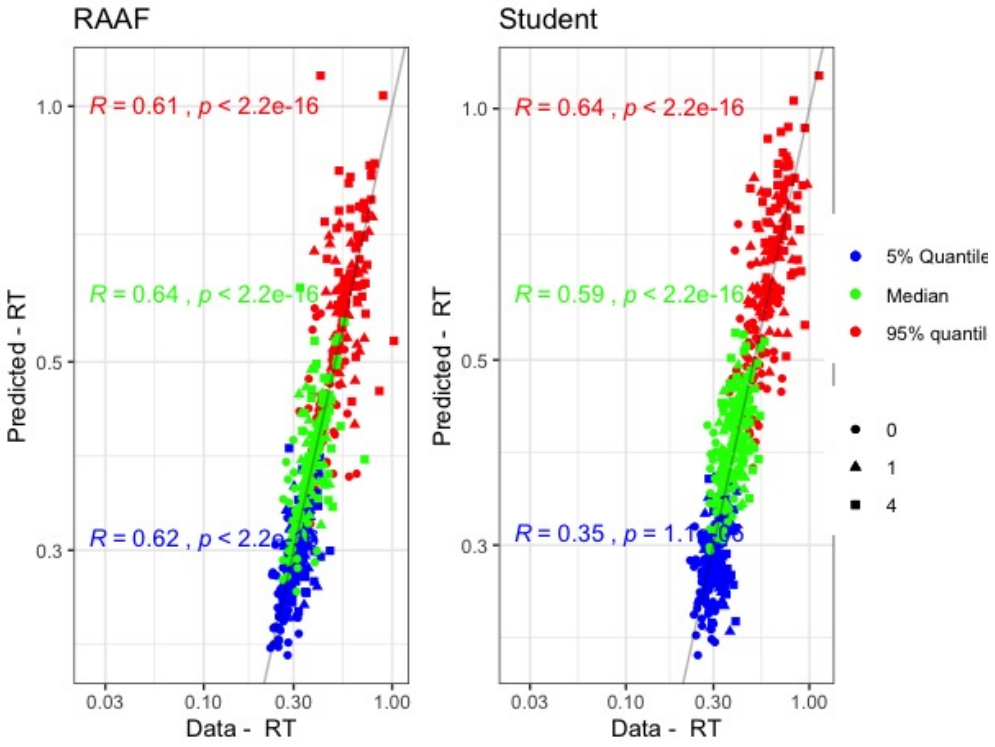


FIGURE B.1: Median, 5% and 95% DRT response times across subjects from posterior predictive data (y-axis) and observed data (x-axis) between groups.

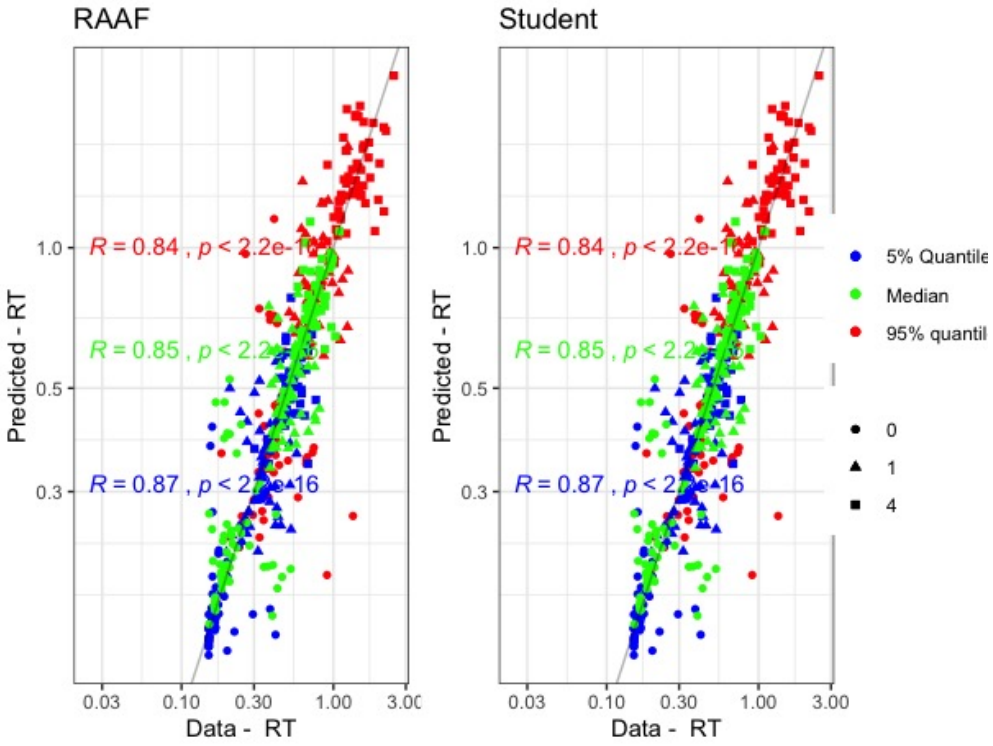


FIGURE B.2: Median, 5% and 95% MOT response times across subjects from posterior predictive data (y-axis) and observed data (x-axis) between groups.

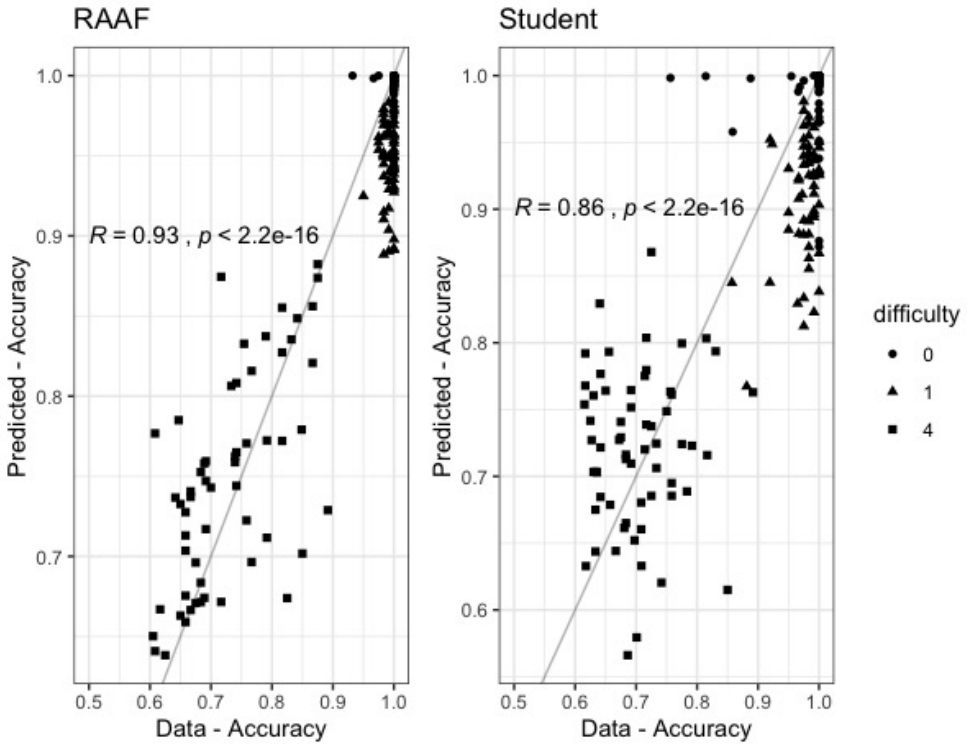


FIGURE B.3: Median MOT accuracy across subjects from posterior predictive data (y-axis) and observed data (x-axis) between groups.

B.2 Further Plots Model Results

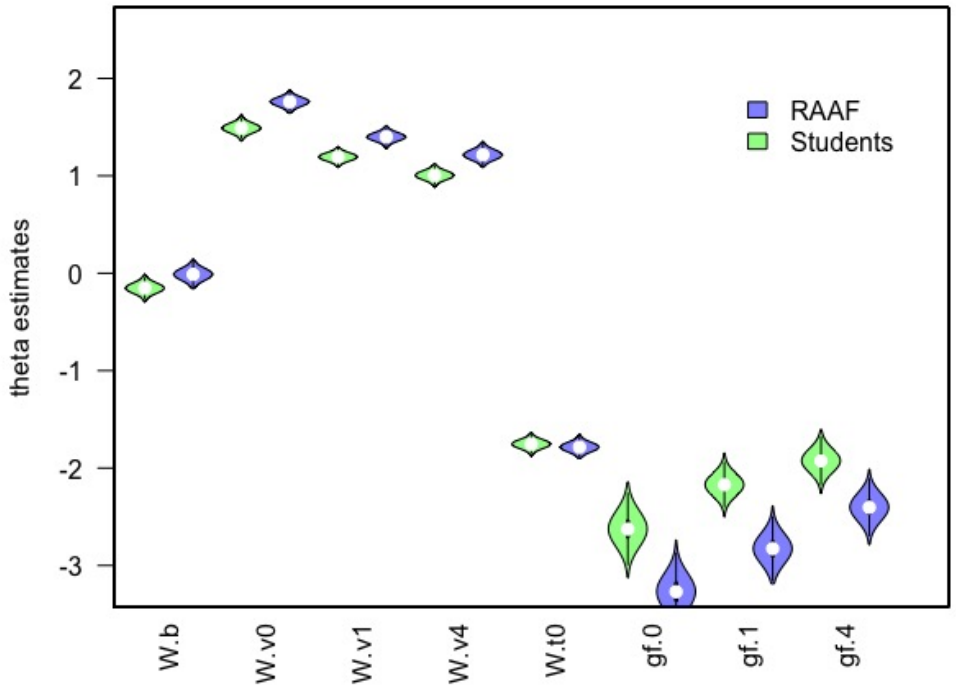


FIGURE B.4: Violin plots for group level parameter estimates from the shifted-Wald across groups. Parameters are shown across the x-axis. Estimates are shown as the log of the estimated value (except for the go-failure parameters, estimated according to a probit estimation).

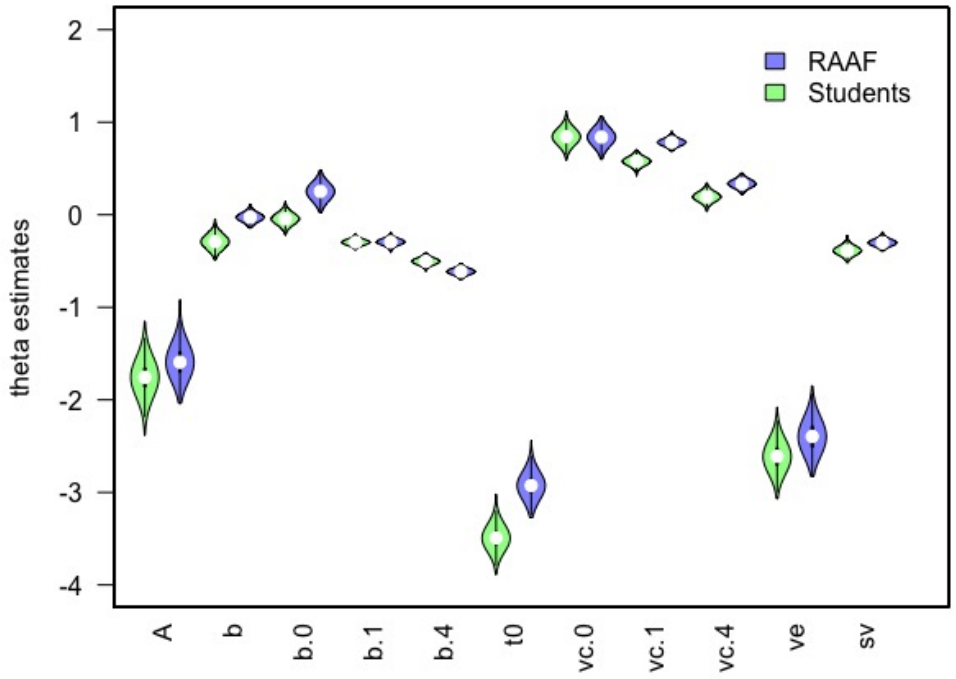


FIGURE B.5: Violin plots for group level parameter estimates from the LBA across groups. Parameters are shown across the x-axis. Estimates are shown as the log of the estimated value.

B.3 Tables

TABLE B.1: Full correlation matrix from Experiment 4 for RAAF data.

	b	$v^{(0)}$	$v^{(1)}$	$v^{(4)}$	T_0	$f^{(0)}$	$f^{(1)}$	$f^{(4)}$	A	b	$b^{(0)}$	$b^{(1)}$	$b^{(4)}$	T_0	$v_c^{(0)}$	$v_c^{(1)}$	$v_c^{(4)}$	v_e	sv
b	1.00	0.50	0.59	0.52	-0.59	-0.14	-0.24	-0.26	0.07	0.12	-0.08	-0.04	-0.02	-0.11	0.05	0.11	0.29	0.07	-0.05
$v^{(0)}$	0.50	1.00	0.70	0.61	-0.35	-0.43	-0.40	-0.39	0.01	0.12	-0.03	-0.13	0.05	0.14	-0.05	0.08	0.29	0.10	-0.03
$v^{(0)}$	0.59	0.70	1.00	0.71	-0.47	-0.41	-0.46	-0.47	-0.07	0.08	-0.02	-0.15	0.02	0.08	-0.05	0.13	0.33	0.11	-0.04
$v^{(0)}$	0.52	0.61	0.71	1.00	-0.43	-0.37	-0.42	-0.46	-0.13	0.02	0.02	-0.12	-0.01	0.03	-0.11	0.09	0.27	0.09	-0.12
T_0	-0.59	-0.35	-0.47	-0.43	1.00	0.04	0.16	0.22	-0.02	-0.08	0.03	0.02	0.15	0.14	-0.04	-0.15	-0.25	-0.07	0.03
$f^{(0)}$	-0.14	-0.43	-0.41	-0.37	0.04	1.00	0.47	0.48	0.04	0.08	0.12	0.09	0.02	-0.13	-0.08	-0.03	-0.17	0.06	0.07
$f^{(1)}$	-0.24	-0.40	-0.46	-0.42	0.16	0.47	1.00	0.57	0.08	0.11	-0.06	0.11	0.04	-0.12	0.13	-0.02	-0.22	-0.01	0.13
$f^{(4)}$	-0.26	-0.39	-0.47	-0.46	0.22	0.48	0.57	1.00	0.11	0.11	-0.11	0.06	0.08	-0.07	0.14	-0.09	-0.17	-0.02	0.08
A	0.07	0.01	-0.07	-0.13	-0.02	0.04	0.08	0.11	1.00	-0.27	-0.25	0.13	0.13	-0.24	0.29	-0.02	-0.08	-0.06	-0.27
b	0.12	0.12	0.08	0.02	-0.08	0.08	0.11	0.11	-0.27	1.00	0.19	-0.11	-0.10	0.13	-0.10	0.16	0.23	0.23	0.37
$b^{(0)}$	-0.08	-0.03	-0.02	0.02	0.03	0.12	-0.06	-0.11	-0.25	0.19	1.00	-0.18	-0.04	0.58	-0.74	0.24	0.28	0.45	-0.00
$b^{(1)}$	-0.04	-0.13	-0.15	-0.12	0.02	0.09	0.11	0.06	0.13	-0.11	-0.18	1.00	0.09	-0.29	0.25	-0.28	-0.18	-0.13	0.05
$b^{(4)}$	-0.02	0.05	0.02	-0.01	0.15	0.02	0.04	0.08	0.13	-0.10	-0.04	0.09	1.00	0.06	-0.01	-0.10	0.01	0.09	-0.11
T_0	-0.11	0.14	0.08	0.03	0.14	-0.13	-0.12	-0.07	-0.24	0.13	0.58	-0.29	0.06	1.00	-0.56	0.34	0.27	0.41	0.01
$v_c^{(0)}$	0.05	-0.05	-0.05	-0.11	-0.04	-0.08	0.13	0.14	0.29	-0.10	-0.74	0.25	-0.01	-0.56	1.00	-0.16	-0.23	-0.39	0.16
$v_c^{(1)}$	0.11	0.08	0.13	0.09	-0.15	-0.03	-0.02	-0.09	-0.02	0.16	0.24	-0.28	-0.10	0.34	-0.16	1.00	0.33	0.36	0.11
$v_c^{(4)}$	0.29	0.29	0.33	0.27	-0.25	-0.17	-0.22	-0.17	-0.08	0.23	0.28	-0.18	0.01	0.27	-0.23	0.33	1.00	0.32	-0.04
v_e	0.07	0.10	0.11	0.09	-0.07	0.06	-0.01	-0.02	-0.06	0.23	0.45	-0.13	0.09	0.41	-0.39	0.36	0.32	1.00	0.08
sv	-0.05	-0.03	-0.04	-0.12	0.03	0.07	0.13	0.08	-0.27	0.37	-0.00	0.05	-0.11	0.01	0.16	0.11	-0.04	0.08	1.00

TABLE B.2: Full correlation matrix from Experiment 4 for Student data.

	b	$v^{(0)}$	$v^{(1)}$	$v^{(4)}$	T_0	$f^{(0)}$	$f^{(1)}$	$f^{(4)}$	A	b	$b^{(0)}$	$b^{(1)}$	$b^{(4)}$	T_0	$v_c^{(0)}$	$v_c^{(1)}$	$v_c^{(4)}$	v_e	sv
b	1.00	0.46	0.50	0.52	-0.56	-0.25	-0.24	-0.18	0.15	0.12	-0.03	-0.09	-0.22	-0.05	0.14	0.23	0.16	0.21	-0.04
$v^{(0)}$	0.46	1.00	0.61	0.61	-0.18	-0.47	-0.43	-0.45	0.15	0.15	0.06	0.05	-0.13	0.09	0.18	0.30	0.20	0.33	0.13
$v^{(0)}$	0.50	0.61	1.00	0.70	-0.28	-0.42	-0.44	-0.46	0.03	0.02	0.06	0.04	-0.20	0.16	0.03	0.23	0.22	0.30	0.03
$v^{(0)}$	0.52	0.61	0.70	1.00	-0.32	-0.40	-0.39	-0.42	0.07	-0.11	-0.05	0.06	-0.05	0.04	0.07	0.08	0.09	0.28	-0.02
T_0	-0.56	-0.18	-0.28	-0.32	1.00	0.04	0.04	-0.06	-0.07	0.04	0.18	0.07	0.06	0.11	-0.19	-0.04	-0.03	-0.07	0.12
$f^{(0)}$	-0.25	-0.47	-0.42	-0.40	0.04	1.00	0.59	0.48	-0.21	-0.12	-0.06	-0.10	0.17	0.15	-0.14	-0.19	-0.17	-0.35	-0.17
$f^{(1)}$	-0.24	-0.43	-0.44	-0.39	0.04	0.59	1.00	0.61	-0.06	-0.04	-0.07	0.03	0.22	-0.04	-0.04	-0.17	-0.12	-0.26	-0.17
$f^{(4)}$	-0.18	-0.45	-0.46	-0.42	-0.06	0.48	0.61	1.00	0.03	-0.07	-0.13	-0.04	0.18	-0.15	0.04	-0.11	-0.13	-0.26	-0.11
A	0.15	0.15	0.03	0.07	-0.07	-0.21	-0.06	0.03	1.00	-0.02	-0.38	-0.10	-0.06	-0.34	0.44	0.29	0.30	0.27	-0.13
b	0.12	0.15	0.02	-0.11	0.04	-0.12	-0.04	-0.07	-0.02	1.00	0.41	-0.19	-0.38	-0.07	-0.02	0.44	0.34	0.12	0.03
$b^{(0)}$	-0.03	0.06	0.06	-0.05	0.18	-0.06	-0.07	-0.13	-0.38	0.41	1.00	-0.11	-0.27	0.23	-0.51	0.25	0.19	-0.01	-0.00
$b^{(1)}$	-0.09	0.05	0.04	0.06	0.07	-0.10	0.03	-0.04	-0.10	-0.19	-0.11	1.00	0.32	-0.01	0.21	-0.30	-0.27	0.02	0.25
$b^{(4)}$	-0.22	-0.13	-0.20	-0.05	0.06	0.17	0.22	0.18	-0.06	-0.38	-0.27	0.32	1.00	-0.03	0.17	-0.38	-0.35	-0.06	0.05
T_0	-0.05	0.09	0.16	0.04	0.11	0.15	-0.04	-0.15	-0.34	-0.07	0.23	-0.01	-0.03	1.00	-0.27	0.04	0.04	-0.02	0.16
$v_c^{(0)}$	0.14	0.18	0.03	0.07	-0.19	-0.14	-0.04	0.04	0.44	-0.02	-0.51	0.21	0.17	-0.27	1.00	0.09	-0.00	0.20	0.17
$v_c^{(1)}$	0.23	0.30	0.23	0.08	-0.04	-0.19	-0.17	-0.11	0.29	0.44	0.25	-0.30	-0.38	0.04	0.09	1.00	0.57	0.32	-0.10
$v_c^{(4)}$	0.16	0.20	0.22	0.09	-0.03	-0.17	-0.12	-0.13	0.30	0.34	0.19	-0.27	-0.35	0.04	-0.00	0.57	1.00	0.34	-0.27
v_e	0.21	0.33	0.30	0.28	-0.07	-0.35	-0.26	-0.26	0.27	0.12	-0.01	0.02	-0.06	-0.02	0.20	0.32	0.34	1.00	-0.06
sv	-0.04	0.13	0.03	-0.02	0.12	-0.17	-0.17	-0.11	-0.13	0.03	-0.00	0.25	0.05	0.16	0.17	-0.10	-0.27	-0.06	1.00

References

Adler, R. F., & Benbunan-Fich, R. (2012). Juggling on a high wire: Multitasking effects on performance. *International Journal of Human-Computer Studies*, *70*(2), 156–168.

Aghajani, H., Garbey, M., & Omurtag, A. (2017). Measuring mental workload with eeg+ fnirs. *Frontiers in human neuroscience*, *11*, 359.

Ahlstrom, U., & Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics*, *36*(7), 623–636.

Anders, R., Alario, F., Van Maanen, L., et al. (2016). The shifted wald distribution for response time data analysis. *Psychological methods*, *21*(3), 309.

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of learning and motivation*, *8*, 47–89.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., ... Craven, P. L. (2007). Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, *78*(5), B231–B244.

Biondi, F. N., Balasingam, B., & Ayare, P. (2020). On the cost of detection response task performance on cognitive load. *Human factors*.

Biondi, F. N., Lohani, M., Hopman, R., Mills, S., Cooper, J. M., & Strayer, D. L. (2018). 80 mph and out-of-the-loop: Effects of real-world semi-automated driving on driver workload and arousal. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 62, pp. 1878–1882).

Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annu. Rev. Psychol.*, *55*, 803–832.

Borst, J. P., Taatgen, N. A., & Van Rijn, H. (2010). The problem state: A cognitive bottleneck in multitasking. *Journal of Experimental Psychology: Learning, memory, and cognition*, *36*(2), 363.

Broadbent, D. E. (2013). *Perception and communication*. Elsevier.

Brock, D., Stroup, J. L., & Ballas, J. A. (2002). Effects of 3d auditory display on dual task performance in a simulated multiscreen watchstation environment. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 46, pp. 1570–1573).

Brown, S., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, *112*, 117-128.

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

Brünken, R., Plass, J. L., & Leutner, D. (2004). Assessment of cognitive load in multimedia learning with dual-task methodology: Auditory load and modality effects. *Instructional Science*, *32*(1-2), 115–132.

Brünken, R., Steinbacher, S., Plass, J. L., & Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Experimental psychology*, *49*(2), 109.

Castro, S. C., Strayer, D. L., Matzke, D., & Heathcote, A. (2019). Cognitive workload measurement and modeling under divided attention. *Journal of experimental psychology: human perception and performance*, *45*(6), 826.

Causse, M., Chua, Z., Peysakhovich, V., Del Campo, N., & Matton, N. (2017). Mental workload and neural efficiency quantified in the prefrontal cortex using fnirs. *Scientific reports*, *7*(1), 1–15.

Causse, M., Peysakhovich, V., & Fabre, E. F. (2016). High working memory load impairs language processing during a simulated piloting task: an erp and pupillometry study. *Frontiers in human neuroscience*, *10*, 240.

Coleman, J. R., Turrill, J., Cooper, J. M., & Strayer, D. L. (2016). Cognitive workload using interactive voice messaging systems. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 60, pp. 1894–1898).

Conti, A. S., Dlugosch, C., & Bengler, K. (2014). The effect of task set instruction on detection response task performance. *Proceedings of the human factors and ergonomics society europe*, 107–117.

Conti, A. S., Dlugosch, C., Vilimek, R., Keinath, A., & Bengler, K. (2012). An assessment of cognitive workload using detection response tasks. *Advances in human aspects of road and rail transport*, 735–743.

Cooper, J. M., Castro, S. C., & Strayer, D. L. (2016). Extending the detection response task to simultaneously measure cognitive and visual task demands. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 60, pp. 1962–1966).

Cooper, J. M., & Strayer, D. L. (2008). Effects of simulator practice and real-world experience on cell-phone—related driver distraction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(6), 893–902.

Dan, A., & Reiner, M. (2017). Eeg-based cognitive load of processing events in 3d virtual worlds is lower than processing events in 2d displays. *International Journal of Psychophysiology*, *122*, 75–84.

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior research methods*, *40*(2), 428–434.

Davidson, M. (2014). Known-groups validity. *Encyclopedia of quality of life and well-being research*, 3481–3482.

de Hollander, G., Forstmann, B. U., & Brown, S. D. (2016). Different ways of linking behavioral and neural data via computational cognitive models. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(2), 101–109.

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional science*, *38*(2), 105–134.

Diederich, A., & Colonius, H. (2004). Bimodal and trimodal multisensory enhancement: Effects of stimulus onset and intensity on reaction time. *Perception & Psychophysics*, *66*, 1388–1404.

Diels, C. (2011). Tactile detection task as a real time cognitive workload measure. In *Contemporary ergonomics and human factors 2011: Proceedings of the international conference on ergonomics & human factors 2011, stoke rochford, lincolnshire, 12-14 april 2011* (p. 183).

Dillon, D. G., Wiecki, T., Pechtel, P., Webb, C., Goer, F., Murray, L., ... others (2015). A computational analysis of flanker interference in depression. *Psychological medicine*,

45(11), 2333–2344.

Donkin, C., Brown, S., Heathcote, A. J., & Wagenmakers, E.-J. (2011). Diffusion versus linear ballistic accumulation: Different models for response time, same conclusions about psychological mechanisms? *Psychonomic Bulletin & Review*, *55*, 140-151.

Drew, T., McCollough, A. W., Horowitz, T. S., & Vogel, E. K. (2009). Attentional enhancement during multiple-object tracking. *Psychonomic Bulletin & Review*, *16*(2), 411–417.

Drews, F. A., Pasupathi, M., & Strayer, D. L. (2008). Passenger and cell phone conversations in simulated driving. *Journal of Experimental Psychology: Applied*, *14*(4), 392.

Eidels, A., Donkin, C., Brown, S. D., & Heathcote, A. (2010). Converging measures of workload capacity. *Psychonomic Bulletin & Review*.

Eidels, A., Townsend, J. T., Hughes, H. C., & Perry, L. A. (2015). Evaluating perceptual integration: Uniting response-time-and accuracy-based methodologies. *Attention, Perception, & Psychophysics*, *77*(2), 659–680.

Engström, J., Åberg, N., Johansson, E., & Hammarbäck, J. (2005). Comparison between visual and tactile signal detection tasks applied to the safety assessment of in-vehicle information systems. University of Iowa.

Engström, J., Larsson, P., & Larsson, C. (2013). Comparison of static and driving simulator venues for the tactile detection response task. In *Proceedings of the seventh international driving symposium on human factors in driver assessment, training, and vehicle design* (pp. 369–375).

Eppler, M. J., & Mengis, J. (2008). The concept of information overload-a review of literature from organization science, accounting, marketing, mis, and related disciplines (2004). In *Kommunikationsmanagement im wandel* (pp. 271–305). Springer.

Evans, N. J. (2020). Same model, different conclusions: An identifiability issue in the linear ballistic accumulator model of decision-making.

Evans, N. J., Rae, B., Bushmakin, M., Rubin, M., & Brown, S. D. (2017). Need for closure is associated with urgency in perceptual decision-making. *Memory & cognition*, *45*(7), 1193–1205.

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

Farrell, S., Ratcliff, R., Cherian, A., & Segraves, M. (2006). Modeling unidimensional categorization in monkeys. *Learning and Behavior*, *34*, 86–101.

Feigh, K. M., Dorneich, M. C., & Hayes, C. C. (2012). Toward a characterization of adaptive systems: A framework for researchers and system designers. *Human Factors*, *54*(6), 1008–1024.

Forstmann, B. U., Anwander, A., Schäfer, A., Neumann, J., Brown, S., Wagenmakers, E.-J., ... Turner, R. (2010). Cortico–striatal connections predict control over speed and accuracy in perceptual decision making. *Proceedings of the National Academy of Science*, *107*, 15916–15920.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., Von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre-sma facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences*, *105*(45), 17538–17542.

Forstmann, B. U., Tittgemeyer, M., Wagenmakers, E.-J., Derrfuss, J., Imperati, D., & Brown, S. (2011). The speed-accuracy tradeoff in the elderly brain: a structural model-based approach. *Journal of Neuroscience*, *31*(47), 17242–17249.

Forstmann, B. U., & Wagenmakers, E.-J. (2015). *An introduction to model–based cognitive neuroscience*. New York: Springer.

Frank, M. J., Scheres, A., & Sherman, S. J. (2007). Understanding decision making deficits in neurological conditions: Insights from models of natural action selection. *Philosophical Transactions of the Royal Society, Series B*, *362*, 1641–1654.

Frey, J., Daniel, M., Castet, J., Hachet, M., & Lotte, F. (2016). Framework for electroencephalography-based evaluation of user experience. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 2283–2294).

Gaetan, S., Dousset, E., Marqueste, T., Bringoux, L., Bourdin, C., Vercher, J.-L., & Besson, P. (2015). Cognitive workload and psychophysiological parameters during multitask activity in helicopter pilots. *Aerospace medicine and human performance*, *86*(12), 1052–1057.

Gawron, V. J. (2019). *Human performance, workload, and situational awareness measures handbook, -2-volume set*. CRC Press.

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., & Zander, T. O. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in neuroscience*, *8*, 385.

Gevins, A., Smith, M. E., Leong, H., McEvoy, L., Whitfield, S., Du, R., & Rush, G. (1998). Monitoring working memory load during computer-based tasks with eeg pattern recognition methods. *Human factors*, *40*(1), 79–91.

Gevins, A., Smith, M. E., McEvoy, L., & Yu, D. (1997). High-resolution eeg mapping of cortical activation related to working memory: effects of task difficulty, type of processing, and practice. *Cerebral cortex (New York, NY: 1991)*, *7*(4), 374–385.

Gilson, R. D. (1995). Special issue preface. *Human Factors*, *37*(1), 3–4.

Gopher, D., Armony, L., & Greenshpan, Y. (2000). Switching tasks and attention policies. *Journal of Experimental Psychology: General*, *129*(3), 308.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? a comparative analysis of six preconceptions about internet questionnaires. *American psychologist*, *59*(2), 93.

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

Grober, E., Buschke, H., Crystal, H., Bang, S., & Dresner, R. (1988). Screening for dementia by memory testing. *Neurology*, *38*(6), 900–900.

Gross, B., Bretschneider-Hagemes, M., Stefan, A., & Rissler, J. (2018). Monitors vs. smart glasses: A study on cognitive workload of digital information systems on forklift trucks. In *International conference on digital human modeling and applications in health, safety, ergonomics and risk management* (pp. 569–578).

Gunawan, D., Hawkins, G. E., Tran, M.-N., Kohn, R., & Brown, S. D. (2020). New estimation approaches for the hierarchical linear ballistic accumulator model. *Journal of Mathematical Psychology*, *96*, 102368.

Haapalainen, E., Kim, S., Forlizzi, J. F., & Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th acm international conference on ubiquitous computing* (pp. 301–310).

Hancock, P. A., & Matthews, G. (2019). Workload and performance: Associations, insensitivities, and dissociations. *Human factors*, *61*(3), 374–392.

Hannula, M., Huttunen, K., Koskelo, J., Laitinen, T., & Leino, T. (2008). Comparison between artificial neural network and multilinear regression models in an evaluation of

cognitive workload in a flight simulator. *Computers in biology and medicine*, *38*(11-12), 1163–1170.

Hardman, D., & Macchi, L. (2004). *Thinking: psychological perspectives on reasoning, judgment and decision making*. John Wiley & Sons.

Haritaoglu, I., Harwood, D., & Davis, L. S. (1998). W 4 s: A real-time system for detecting and tracking people in 2 1/2d. In *European conference on computer vision* (pp. 877–892).

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, pp. 904–908).

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier.

Hawkins, G., & Heathcote, A. (2019). Racing against the clock: Evidence-based vs. time-based decisions.

Hawkins, G. E., Marley, A., Heathcote, A., Flynn, T. N., Louviere, J. J., & Brown, S. D. (2014). Integrating cognitive process and descriptive models of attitudes and preferences. *Cognitive science*, *38*(4), 701–735.

Hays, R. T., Jacobs, J. W., Prince, C., & Salas, E. (1992). Flight simulator training effectiveness: A meta-analysis. *Military psychology*, *4*(2), 63–74.

Heathcote, A. (2004). Fitting Wald and ex–Wald distributions to response time data: An example using functions for the S–PLUS package. *Behavior Research Methods, Instruments, & Computers*, *36*, 678–694.

Heathcote, A., Suraev, A., Curley, S., Gong, Q., Love, J., & Michie, P. T. (2015). Decision processes and the slowing of simple choices in schizophrenia. *Journal of Abnormal Psychology*, *124*(4), 961.

Hedge, C., Vivian-Griffiths, S., Powell, G., Bompas, A., & Sumner, P. (2019). Slow and steady? Strategic adjustments in response caution are moderately reliable and correlate across tasks. *Consciousness and Cognition*, *75*, 102797.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., & Schultz, T. (2014). Mental workload during n-back task???quantified in the prefrontal cortex using fnirs. *Frontiers in human neuroscience*, *7*, 935.

Hicks, J. L., Althoff, T., Kuhar, P., Bostjancic, B., King, A. C., Leskovec, J., ... others (2019). Best practices for analyzing large-scale health data from wearables and smartphone apps. *NPJ digital medicine*, *2*(1), 1–12.

Ho, T. C., Yang, G., Wu, J., Cassey, P., Brown, S. D., Hoang, N., ... others (2014). Functional connectivity of negative emotional processing in adolescent depression. *Journal of affective disorders*, *155*, 65–74.

Howard, Z. L., Evans, N. E., Innes, R. J., Brown, S. D., & Eidels, A. (2020). How is multi-tasking different from increased difficulty? *Psychonomic Bulletin & Review*.

Howard, Z. L., Innes, R. J., Brown, S. D., & Eidels, A. (2018). *Cognitive workload and analysis of flight path data* (Tech. Rep.). Callaghan, NSW, Australia: University of Newcastle.

Huang, A., & Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, *8*(2), 439-452.

Hughes, A. M., Hancock, G. M., Marlow, S. L., Stowers, K., & Salas, E. (2019). Cardiac measures of cognitive workload: a meta-analysis. *Human factors*, *61*(3), 393–414.

Huttunen, K., Keränen, H., Väyrynen, E., Pääkkönen, R., & Leino, T. (2011). Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights.

Applied ergonomics, *42*(2), 348–357.

Innes, R. J., Evans, N. J., Howard, Z. L., Eidels, A., & Brown, S. D. (2020). A broader application of the detection response task to cognitive tasks and online environments. *Human factors*. doi: https://doi.org/10.1177/0018720820936800

Innes, R. J., Howard, Z. L., Eidels, A., & Brown, S. D. (2018). *Cognitive workload measurement and analysis* (Tech. Rep.). Callaghan, NSW, Australia: University of Newcastle.

Innes, R. J., Howard, Z. L., Thorpe, A., Eidels, A., & Brown, S. D. (2020). The effects of increased visual information on cognitive workload in a helicopter simulator. *Human factors*. doi: https://doi.org/10.1177/0018720820945409

Innes, R. J., & Kuhne, C. L. (2020). An lba account of decisions in the multiple object tracking task. *The Quantitative Methods for Psychology*, *16*, 175–191.

ISO:17488. (2016). *Road vehicles–transport information and control systems–detection-response task (drt) for assessing attentional effects of cognitive load in driving.* International Organization for Standardization Geneva, Switzerland.

Jaeggi, S. M., Buschkuehl, M., Etienne, A., Ozdoba, C., Perrig, W. J., & Nirkko, A. C. (2007). On how high performers keep cool brains in situations of cognitive overload. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(2), 75–89.

JASP Team. (2019). *JASP (Version 0.11.0)[Computer software].* Retrieved from `https://jasp-stats.org/`

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Johnson, D. M., & Wiles, J. (2003). Effective affective user interface design in games. *Ergonomics*, *46*(13/14), 1332-1345.

Kahneman, D. (1973). *Attention and effort*. Citeseer.

Kantowitz, B. H., & Casper, P. A. (2017). Human workload in aviation. In *Human error in aviation* (pp. 123–153). Routledge.

Karşılar, H., Simen, P., Papadakis, S., & Balcı, F. (2014). Speed accuracy trade-off under response deadlines. *Frontiers in neuroscience*, *8*, 248.

Klingner, J. M. (2010). *Measuring cognitive load during visual tasks by combining pupillometry and eye tracking* (Unpublished doctoral dissertation). Stanford University Palo Alto, CA.

Koller, D., Weber, J., & Malik, J. (1994). Robust multiple car tracking with occlusion reasoning. In *European conference on computer vision* (pp. 189–196).

Krantz, J. H., & Dalal, R. (2000). Validity of web-based psychological research. In *Psychological experiments on the internet* (pp. 35–60). Elsevier.

Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., & Krejtz, I. (2018). Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PloS one*, *13*(9).

Krueger, G. P., Armstrong, R. N., & Cisco, R. R. (1985). Aviator performance in week-long extended flight operations in a helicopter simulator. *Behavior Research Methods, Instruments, & Computers*, *17*(1), 68–74.

Lavie, N. (2010). Attention, distraction, and cognitive control under load. *Current directions in psychological science*, *19*(3), 143–148.

Lee, J. D., Young, K. L., & Regan, M. A. (2008). Defining driver distraction. *Driver distraction: Theory, effects, and mitigation*, *13*(4), 31–40.

Lerche, V., & Voss, A. (2017). Retest reliability of the parameters of the Ratcliff diffusion model. *Psychological Research*, *81*(3), 629–652.

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., Jooste, L., Li, Y. R., Ballard, T., Huf, S., Lipp, O. V., & Visser, T. A. (2018). Using situation awareness and workload to predict performance in submarine track management: A multilevel approach. *Human factors*, *60*(7), 978–991.

Lohani, M., Payne, B. R., & Strayer, D. L. (2019). A review of psychophysiological measures to assess cognitive states in real-world driving. *Frontiers in human neuroscience*, *13*.

Marek, T., Karwowski, W., & Rice, V. (2010). Cognitive workload assessment of air traffic controllers using optical brain imaging sensors. In *Advances in understanding human performance* (pp. 36–46). CRC Press.

Matthews, G., Reinerman-Jones, L. E., Barber, D. J., & Abich IV, J. (2015). The psychometrics of mental workload: Multiple measures are sensitive but divergent. *Human factors*, *57*(1), 125–143.

Matzke, D., Hughes, M., Badcock, J. C., Michie, P., & Heathcote, A. (2017). Failures of cognitive control or attention? The case of stop-signal deficits in schizophrenia. *Attention, Perception, & Psychophysics*, *79*(4), 1078–1086.

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*, 798–817.

Mayhew, D. J. (1999). *The usability engineering lifecycle: a practitioner's handbook for user interface design*. Morgan Kaufmann.

McKerral, A., Boyce, N., & Pammer, K. (2019). Supervising the self-driving car: situation awareness and fatigue during automated driving. In *Proceedings of the 11th international conference on automotive user interfaces and interactive vehicular applications: Adjunct proceedings* (pp. 315–320).

Medeiros-Ward, N., Watson, J. M., & Strayer, D. L. (2015). On supertaskers and the neural basis of efficient multitasking. *Psychonomic Bulletin & Review*, *22*(3), 876–883.

Mehler, B., Reimer, B., Coughlin, J., & Dusek, J. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record: Journal of the Transportation Research Board*(2138), 6–12.

Mehler, B., Reimer, B., & Wang, Y. (2011). A comparison of heart rate and heart rate variability indices in distinguishing single-task driving and driving under secondary cognitive workload. In *Proceedings of the sixth international driving symposium on human factors in driver assessment, training and vehicle design.* University of Iowa.

Merat, N., & Jamson, A. H. (2008). The effect of stimulus modality on signal detection: Implications for assessing the safety of in-vehicle technology. *Human factors*, *50*(1), 145–158.

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part i. basic mechanisms. *Psychological review*, *104*(1), 3.

Meyerson, P., & Tryon, W. W. (2003). Validating internet research: A test of the psychometric equivalence of internet and in-person samples. *Behavior Research Methods, Instruments, & Computers*, *35*(4), 614–620.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

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

Morey, R. D., & Rouder, J. N. (2013). BayesFactor: Computation of Bayes factors for simple designs [Computer software manual]. (R package version 0.9.4)

Motti, V. G., & Caine, K. (2015). Users' privacy concerns about wearables. In *International conference on financial cryptography and data security* (pp. 231–244).

Mühl, C., Jeunet, C., & Lotte, F. (2014). Eeg-based workload estimation across affective contexts. *Frontiers in neuroscience*, *8*, 114.

Münsterer, T., Schafhitzel, T., Strobel, M., Völschow, P., Klasen, S., & Eisenkeil, F. (2014). Sensor-enhanced 3d conformal cueing for safe and reliable hc operation in dve in all flight phases. In *Degraded visual environments: Enhanced, synthetic, and external vision solutions 2014* (Vol. 9087, p. 90870I).

Nelson, W. (1988). *Use of circular error probability in target detection*. Defense Technical Information Center.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar–based random walk model of speeded classification. *Psychological Review*, *104*, 266–300.

Nourbakhsh, N., Wang, Y., & Chen, F. (2013). Gsr and blink features for cognitive load classification. In *Ifip conference on human-computer interaction* (pp. 159–166).

Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. A. (2012). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th australian computer-human interaction conference* (pp. 420–423).

Osth, A. F., Jansson, A., Dennis, S., & Heathcote, A. (2018). Modeling the dynamics of recognition memory testing with an integrated model of retrieval and decision making. *Cognitive psychology*, *104*, 106–142.

Palestro, J. J., Bahg, G., Sederberg, P. B., Lu, Z.-L., Steyvers, M., & Turner, B. M. (2018). A tutorial on joint models of neural and behavioral measures of cognition. *Journal of Mathematical Psychology*, *84*, 20–48.

Palmer, J. (1990). Attentional limits on the perception and memory of visual information. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(2), 332.

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. (1994). Dual-task interference in simple tasks: data and theory. *Psychological bulletin*, *116*(2), 220.

Pashler, H. (1998). *The psychology of attention*. MIT press.

Pashler, H. (2000). 12 task switching and multitask performance. *Control of cognitive processes*, 277.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Rabbitt, P. (1979). How old and young subjects monitor and control responses for accuracy and speed. *British Journal of Psychology*, *70*, 305–311.

Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. D. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1226.

Randall, D. M., & Fernandes, M. F. (1991). The social desirability response bias in ethics research. *Journal of business ethics*, *10*(11), 805–817.

Rasmussen, C., & Hager, G. D. (1998). Joint probabilistic techniques for tracking multi-part objects. In *Proceedings. 1998 ieee computer society conference on computer vision and pattern recognition (cat. no. 98cb36231)* (pp. 16–21).

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological review*, *111*(1), 159.

Ratcliff, R., Hasegawa, Y. T., Hasegawa, Y. P., Smith, P. L., & Segraves, M. A. (2007).

Dual diffusion model for single–cell recording data from the superior colliculus in a brightness–discrimination task. *Journal of Neurophysiology*, *97*, 1756–1774.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two–choice letter identification. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 127–140.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*, 260–281.

Ratcliff, R., & Strayer, D. L. (2014). Modeling simple driving tasks with a one-boundary diffusion model. *Psychonomic bulletin & review*, *21*(3), 577–589.

Ratcliff, R., Thapar, A., & McKoon, G. (2007). Application of the diffusion model to two-choice tasks for adults 75-90 years old. *Psychology and Aging*, *22*, 56–66.

Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and iq in two-choice tasks. *Cognitive psychology*, *60*(3), 127–157.

Ratcliff, R., Thapar, A., & McKoon, G. (2011). Effects of aging and iq on item and associative memory. *Journal of Experimental Psychology: General*, *140*(3), 464.

Reimer, B., & Mehler, B. (2011). The impact of cognitive workload on physiological arousal in young adult drivers: a field study and simulation validation. *Ergonomics*, *54*(10), 932–942.

Roenker, D. L., Cissell, G. M., Ball, K. K., Wadley, V. G., & Edwards, J. D. (2003). Speed-of-processing and driving simulator training result in improved driving performance. *Human factors*, *45*(2), 218–233.

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological review*, *115*(1), 101.

Schacter, D. L. (1999). The seven sins of memory: insights from psychology and cognitive neuroscience. *American psychologist*, *54*(3), 182.

Schmidt, E., Decke, R., Rasshofer, R., & Bullinger, A. C. (2017). Psychophysiological responses to short-term cooling during a simulated monotonous driving task. *Applied ergonomics*, *62*, 9–18.

Selcon, S. J., Taylor, R. M., & Koritsas, E. (1991). Workload or situational awareness?: Tlx vs. sart for aerospace systems design evaluation. In *Proceedings of the human factors society annual meeting* (Vol. 35, pp. 62–66).

Shi, Y., Ruiz, N., Taib, R., Choi, E., & Chen, F. (2007). Galvanic skin response (gsr) as an index of cognitive load. In *Chi'07 extended abstracts on human factors in computing systems* (pp. 2651–2656).

Song, M., Kang, K. Y., Timakum, T., & Zhang, X. (2020). Examining influential factors for acknowledgements classification using supervised learning. *PloS one*, *15*(2), e0228928.

Stanton, N. A., Chambers, P. R., & Piggott, J. (2001). Situational awareness and safety. *Safety science*, *39*(3), 189–204.

Stojmenova, K., Jakus, G., & Sodnik, J. (2017). Sensitivity evaluation of the visual, tactile, and auditory detection response task method while driving. *Traffic injury prevention*, *18*(4), 431–436.

Stojmenova, K., & Sodnik, J. (2018). Detection-response task—uses and limitations. *Sensors*, *18*(2), 594.

Strayer, D. L., Cooper, J. M., McCarty, M. M., Getty, D. J., Wheatley, C. L., Motzkus, C. J., ... Horrey, W. J. (2019). Visual and cognitive demands of carplay, android auto, and five native infotainment systems. *Human Factors*, *61*(8), 1371–1386.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J., Medeiros-Ward, N., & Biondi, F. (2013). Measuring cognitive distraction in the automobile. AAA Foundation for Traffic Safety.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., & Hopman, R. J. (2015). Measuring cognitive distraction in the automobile iii: A comparison of ten 2015 in-vehicle information systems..

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., & Hopman, R. J. (2016). Talking to your car can drive you to distraction. *Cognitive research: principles and implications*, *1*(1), 16.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., & Hopman, R. J. (2017). The smartphone and the driver's cognitive workload: A comparison of apple, google, and microsoft's intelligent personal assistants. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *71*(2), 93.

Strayer, D. L., & Drew, F. A. (2004). Profiles in driver distraction: Effects of cell phone conversations on younger and older drivers. *Human factors*, *46*(4), 640–649.

Strayer, D. L., Drews, F. A., & Crouch, D. J. (2006). A comparison of the cell phone driver and the drunk driver. *Human factors: The journal of the human factors and ergonomics society*, *48*(2), 381–391.

Strayer, D. L., Drews, F. A., & Johnston, W. A. (2003). Cell phone-induced failures of visual attention during simulated driving. *Journal of experimental psychology: Applied*, *9*(1), 23.

Strayer, D. L., & Johnston, W. A. (2001). Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological science*, *12*(6), 462–466.

Strayer, D. L., Turrill, J., Cooper, J. M., Coleman, J. R., Medeiros-Ward, N., & Biondi, F. (2015). Assessing cognitive distraction in the automobile. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(8), 1300–1324.

Strayer, D. L., Watson, J. M., & Drews, F. A. (2011). Cognitive distraction while multitasking in the automobile. *Psychology of Learning and Motivation – Advances in Research and Theory*, *54*, 29.

Svensson, E., Angelborg-Thanderez, M., Sjöberg, L., & Olsson, S. (1997). Information complexity-mental workload and performance in combat aircraft. *Ergonomics*, *40*(3), 362–380.

Thierer, A. D. (2015). The internet of things and wearable technology: Addressing privacy and security concerns without derailing innovation. *Adam Thierer, The Internet of Things and Wearable Technology: Addressing Privacy and Security Concerns without Derailing Innovation*, *21*.

Thorpe, A., Innes, R., Townsend, J., Heath, R., Nesbitt, K., & Eidels, A. (2020). Assessing cross-modal interference in the detection response task. *Journal of Mathematical Psychology*, *98*, 102390.

Thorpe, A., Nesbitt, K., & Eidels, A. (2019). Assessing game interface workload and usability: A cognitive science perspective. In *Proceedings of the australasian computer science week multiconference* (p. 44).

Tillman, G., Strayer, D. L., Eidels, A., & Heathcote, A. (2017). Modeling cognitive load effects of conversation between a passenger and driver. *Attention, Perception, & Psychophysics*, 1–9.

Townsend, J. T., & Eidels, A. (2011). Workload capacity spaces: A unified methodology for response time measures of efficiency as workload is varied. *Psychonomic Bulletin*

& Review, *18*(4), 659–681.

Townsend, J. T., & Wenger, M. J. (2004a). The serial-parallel dilemma: A case study in a linkage of theory and method. *Psychonomic Bulletin & Review*, *11*(3), 391–418.

Townsend, J. T., & Wenger, M. J. (2004b). A theory of interactive parallel processing: new capacity measures and predictions for a response time inequality series. *Psychological review*, *111*(4), 1003.

Treisman, A., & Geffen, G. (1967). Selective attention: Perception or response? *Quarterly Journal of Experimental Psychology*, *19*(1), 1–17.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97–136.

Tsang, P. S., & Vidulich, M. A. (2006). Mental workload and situation awareness. John Wiley & Sons Inc.

Turner, B. M., Forstmann, B. U., Steyvers, M., et al. (2019). *Joint models of neural and behavioral data*. Springer.

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193–206.

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

van den Bergh, D., van Doorn, J., Marsman, M., Draws, T., van Kesteren, E.-J., Derks, K., ... others (2020). A tutorial on conducting and interpreting a bayesian anova in jasp. *LAnnee psychologique*, *120*(1), 73–96.

van Maanen, L., Brown, S. D., Eichele, T., Wagenmakers, E.-J., Ho, T., Serences, J., & Forstmann, B. U. (2011). Neural correlates of trial-to-trial fluctuations in response caution. *Journal of Neuroscience*, *31*(48), 17488–17495.

van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2012). A diffusion model decomposition of the effects of alcohol on perceptual decision making. *Psychopharmacology*, *219*(4), 1017–1025.

Van Winsum, W., Herland, L., & Martens, M. (1999). *The effects of speech versus tactile driver support messages on workload, driver behaviour and user acceptance*. TNO Human Factors Research Institute.

Vashitz, G., Shinar, D., & Blum, Y. (2008). In-vehicle information systems to improve traffic safety in road tunnels. *Transportation Research Part F: Traffic Psychology and Behaviour*, *11*(1), 61–74.

Vidulich, M. A., & Wickens, C. D. (1986). Causes of dissociation between subjective workload measures and performance: Caveats for the use of subjective assessments. *Applied Ergonomics*, *17*(4), 291–296.

Wagenmakers, E.-J., Lee, M. D., Lodewyckx, T., & Iverson, G. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York: Springer Verlag.

Wall, L., Gunawan, D., Brown, S. D., Tran, M.-N., Kohn, R., & Hawkins, G. E. (2019). Identifying relationships between cognitive processes across tasks, contexts, and time. *arXiv preprint arXiv:1910.07185*.

Watson, J. M., & Strayer, D. (2010). Supertaskers: Profiles in extraordinary multitasking ability. *Psychonomic Bulletin & Review*, *17*, 479–485.

Weigard, A., & Huang-Pollock, C. (2014). A diffusion modeling approach to understanding contextual cueing effects in children with ADHD. *Journal of Child Psychology and*

Psychiatry, *55*(12), 1336–1344.

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

Wickens, C. D. (2002b). Situation awareness and workload in aviation. *Current directions in psychological science*, *11*(4), 128–133.

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

Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press, USA.

Wilson, G. F. (2002). An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology*, *12*(1), 3–18.

Xie, F., Wang, Q., Jin, X., Liao, Y., Zheng, S., Li, L., ... Liu, Z. (2016). Evaluation of the crew workload to quantify typical mission profile special vehicles. In *International conference on man-machine-environment system engineering* (pp. 149–158).

Young, R. A., Hsieh, L., & Seaman, S. (2013). The tactile detection response task: preliminary validation for measuring the attentional effects of cognitive load. In *Proceedings of the seventh international driving symposium on human factors in driver assessment, training, and vehicle design* (pp. 71–77).

Zimmermann, M., Gestwa, M., König, C., Wolfram, J., Klasen, S., & Lederle, A. (2019). First results of lidar-aided helicopter approaches during nato dve-mitigation trials. *CEAS Aeronautical Journal*, 1–16.