Design recommendations for managing mental workload transitions in multitasking environments: Evidence from eye tracking and growth curve modeling

A

Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Shannon P. Devlin

August 2021

APPROVAL SHEET

This

Dissertation

is submitted in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

Author: Shannon P. Devlin

This Dissertation has been read and approved by the examing committee:

Advisor: Sara L. Riggs

Advisor:

Committee Member: Gregory J. Gerling

Committee Member: Lu Feng

Committee Member: Homa Alemzadeh

Committee Member: Joseph T. Coyne

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

August 2021

For the Doylestown Devlins + Nick

ACKNOWLEDGMENTS

It took a village to make this dissertation possible, so I would like to take the time to express my gratitude. First and foremost, thank you to my advisor, Dr. Sara Lu Riggs. Your consistently spot-on guidance and unwavering confidence in me has been essential these past five years. You inspire me to tackle complex and challenging problems in ways I never thought was possible. You have advised me on so much more than research, ranging from professional advice to the way you lead and carried yourself each day. I feel so incredibly lucky you took me under your wing as a naïve undergraduate six years ago. I am sorely going to miss our weekly meetings, but I am so excited to watch you and the lab grow and flourish.

I would like to thank the members of my committee for providing me guidance and assistance during this process: Dr. Greg Gerling, Dr. Homa Alemzadeh, Dr. Lu Feng, and Dr. Joseph Coyne. I know my time at UVA was brief, but your fresh insights took this research to "the next level". Thank you for your patience and diligence in working with me as I navigated uncharted territory and challenging me to pursue greater endeavors for this research. Thank you for working with me so flexibly and effortlessly in these unprecedented times. I would also like to thank the UVA StatsLab, specifically Clay Ford, and Dr. Lesa Hoffman at University of Iowa for their generosity in guiding me through the statistical analysis and modeling.

Thank you to the National Science Foundation and the Command Decision Making/Science of AI program at the Office of Naval Research Lab for providing the funds and accessibility to conduct this research. Thank you to the Warfighter Applied Cognition and Technology Laboratory at Naval Research Laboratory for the opportunity to learn about eye tracking, providing much-needed perspective on my research, and the ability to collect data that would not have been possible otherwise. Working with you all has truly been career defining and I cannot wait to see what the future holds!

Thank you to the faculty and staff at Clemson University for your assistance in this research effort, ranging from research guidance to logistical support, and your graciousness as I emigrated to UVA. Thank you to all my coauthors: Dr. Nadine Marie Moacdieh, Hussein Jundi, Dr. Christopher D. Wickens, for helping me navigate the publication process. Thank you to the undergraduates of Riggs Lab: James Hatfield, Aakash Bhagat, Jake Flynn, Jennifer Byham, Samuel Smith, Dustin Nguyen, Jad Atweh, and Mohamad El Iskandarani. All of you have been a tremendous help with programming, experimental design, and data analysis and for teaching me on how to improve my management and leadership skills. As for Riggs Lab graduate students, thank you Scott Betza for your initial assistance with testing and validating the UAV testbed in its earliest stages and for your assistance in landing the DON Pathways internship. Logan Clark, thank you for bringing in a sharp and fresh perspective to Riggs Lab while being a source of humor and mind-bending discussion topics. Thank you, Jawad Alami, for your incredible programming skills and going the extra mile to make the second half of this dissertation possible – I feel so lucky to know someone who is so genuinely generous. Thank you, Dr. Kylie Gomes, for being the best "oldest child" a lab could ever have – your wisdom, guidance, unpretentiousness, and kindness towards me and all lab members set the tone for the lab to be an exciting, innovative, and inclusive space for research.

Thank you to my friends for providing support and life-long memories during this time. I loved getting to know a whole new Clemson crew as a graduate student and traveling with you all to our annual conference each year. Particular thanks to Dr. Katie Jurewicz and Dr. Kylie Gomes who assisted in all aspects of graduate school, leading to a remarkable friendship I hope to have for decades to come. Thank you to my Manning 9th floor ladies for visiting me every fall, providing so much fun, laughter and light, and for being there for me no matter what. To my friends in DC, you made my summer internships unforgettable and were great cheerleaders from afar. I am so excited to continue our friendships in person sometime soon (hopefully).

Thank you to my sisters Kaitlyn and Moira for being my biggest fans from the start. Thank you for barging in to my room unannounced to provide reassurance and comic relief. Thank you for proofreading my papers, hour-long phone calls, and for walking the dogs when it was my turn so I could focus on this dissertation. Thank you for having faith in me, even when I had lost total faith in myself. Mom and Dad, thank you for being the most supportive parents I have ever met. Thank you for helping me with my annual moves in the hottest days of the summer and then ultimately letting me crash at your house for the past 18 months. Dad, thank you for the motivational phone calls when I was too many miles away and for being undeservedly proud in me. Your work ethic inspired me to cross the finish line and I will be proud if I have half of your work ethic in my career. Mom thank you for encouraging me to chase my dreams even in the midst of setbacks. Thank you for doing quite literally everything in your power to help me during all of my so that I could solely focus on the next milestone ahead. Thank you for never holding back your opinion, but always offering unyielding support.

Lastly, but certainly not least, thank you to my better half: Nick. Thank you for being my rock, my sounding board, my cheerleader, my source of fun, love, and laughter, and my teammate. Thank you for knowing how and when to push me to be better – I am always in awe of your ability to know what I need before I do. Your perspective and advice has been irreplaceable throughout the past five years. I hope this dissertation makes you proud. I cannot wait to be doctorates together and take the world by storm.

TABLE OF CONTENTS

List of Tables

List of Figures

ABSTRACT

Complex and dynamic environments including military operations, healthcare, aviation, and driving require operators to seamlessly manage continuous shifts between levels of mental workload, which is known as a *workload transition*. Even though they are expected, there has been limited work examining workload transitions. Currently there is no single theoretical explanation able to unite the findings of workload transition research. For example, there has been limited work examining the effect of transition rate, i.e., the speed at which workload transitions, multiple transitions, multitasking environments, and with context-relevant populations. This limits the ability to provide general design guidance for environments experiencing workload transitions.

One promising way to address the current research gaps is to study visual attention allocation patterns of an individual experiencing a workload transition in real-time. Features inherent to dynamic domains are not often included in workload transition research, which hinders its generalizability. Eye tracking is an increasingly accessible and reliable method to capture the visual attention allocation patterns of a person, which provides the ability to quantify how workload transitions impact the person's mental resources and performance over time.

This dissertation attempts to bridge some of the gaps in the workload transition literature by examining the effect different transition rates have on multitasking performance, performance trends over time, and visual attention allocation patterns within an Unmanned Aerial Vehicle (UAV) command and control environment. The findings add to the workload transition theory and provide design guidance.

CHAPTER 1

Introduction

Dynamic domains like the military (Huey & Wickens, 1993), driving (Morgan & Hancock, 2011), air traffic control (Edwards et al., 2012), and emergency response (Parush & Rustanjaja, 2013) require their operators to seamlessly transition between varying levels of task demands (i.e., workload). Previous research primarily focuses on understanding the effects of low or high workload (Cain, 2007; Grier et al., 2008; Wickens, 2008). However, most dynamic domains do not remain at one constant level of workload, rather they tend to shift between low and high workload, and this can have serious consequences (Williams, 2006). For example, when the Apollo 12 spacecraft was struck by lightning, “all the [alarm] lights came on” the telemetry stream, drastically increasing workload in a short amount of time (Murray & Cox, 1989, p. 374). Although research on workload has received a great deal of attention (see review in Hancock & Matthews, 2019), understanding the consequences of when it changes—i.e., *workload transitions*—has received less to date (Cox-Fuenzalida, 2007; Huey & Wickens, 1993; Prytz & Scerbo, 2015).

Across the workload transition literature, the observed performance of the operator, i.e., the speed and accuracy with which he/she attends to tasks, is not consistent. Performance has been found to be better, worse, or no different than constant workload, depending on the study (detailed review in Bowers, 2013). As for how it trends over time, performance can follow one of these same outcomes or even switch between bettering and worsening (Gluckman et al., 1993). Some research has focused on explaining why performance may be inconsistent (e.g., Ungar et al., 2005), but no single, unifying theory has been developed. Rather, several theoretical explanations exist (Cox-Fuenzalida, 2007; Prytz & Scerbo, 2015; Jansen et al., 2016), but there is uncertainty in their applicability. Even amongst the official calls for research (Huey & Wickens, 1993) and waves of renewed interest (e.g., Cox-Fuenzalida, 2007, Bowers et al., 2014), research rarely discusses how the results add to the theory of workload transition performance, which limits the design guidance on how to properly account for changes in workload.

One promising way to address the present research gaps of workload transitions is to use real-time, physiological data. For example, eye tracking technology can track a person's real-time point of gaze when interacting with a display (Holmqvist et al., 2011; Poole & Ball, 2006). Eye tracking can be used to capture visual attention allocation patterns, i.e., the spatial and temporal properties of overt, selective attention in the visual channel (Moon et al., 2019). Previous research has consistently found these patterns to be a psychophysiological measure, i.e., a biological response due to the mental activity (Gaillard & Kramer, 2009). Research has seldomly examined how an individual's visual attention allocation patterns influence performance when workload transitions, even though the visual channel is inundated with information (Woods et al., 2002). The goal of this research is to see whether eye tracking can shed light on operator's performance during workload transitions in complex, dynamic, realistic work environments. Specifically, it thoroughly explores multitasking performance trends and scan-based eye tracking metrics during different workload transition rates with aggregate and longitudinal analysis methods. The work of this dissertation is among the first to examine whether eye tracking can predict performance when workload transitions over time.

This research occurs in the domain of military unmanned aerial vehicle (UAV) command and control. The United States Department of Defense has expressed the desire to enhance the human-machine collaboration between operators and UAVs, particularly via the design of the visual display (Department of Defense, 2018, p. 29). Currently, it takes several operators to control a single UAV, but the Department of Defense wants one operator to control several UAVs in the future (Cummings et al., 2007; Department of Defense, 2018, p. 20; Goodrich & Cummings, 2015; Sibley et al., 2015). The design of the display is essential to achieving this mission given most, if not all, the pertinent information is visually presented, which differs from manned flight environments (McCarley & Wickens, 2004; Hobbs & Shivley, 2014). The design of the display is also essential for successful human-machine interaction, which is prevalent in UAV command and control (Baker & Keebler, 2017; Cummings et al., 2019). The Department of Defense also wants to diversify the UAV's utility by deploying it in different theatres including combat, surveillance, and maintenance (Freedberg, 2021). The combination of the aforementioned will likely increase the severity and likelihood of workload transitions, which have led to UAV mishaps in the past (Hobbs & Shivley, 2014; Sibley et al., 2015; Williams, 2006). Given the current system requires operators to monitor several visual displays simultaneously, studying visual attention allocation patterns via eye tracking is a promising means to inform the design of future displays (Abich et al., 2017, p. 804; Giese et al., 2013). Eye tracking may also be a tool implemented in UAV command and control, as a part of the greater investment in engineering solutions for current human-machine interaction challenges (Baker & Keebler, 2017; Department of Defense, 2017, p. 20; Giese et al., 2013). This dissertation focuses on aspects of workload transitions relevant to UAV command and control:

- The impact transition rate, i.e., the speed at which workload changes from one load to another has on multitasking performance and performance over time.
- The ability of eye tracking data to explain and predict the observed performance

trends of workload transitions.

- The empirical development of the current theoretical explanations as a means to inform future display design for dynamic environments.

Workload and workload transitions

Unlike workload transitions, there are several definitions in the literature for workload. (Cain, 2007; Van Acker, Parmentier, Vlerick, & Saldien, 2018; Young, Brookhuis, Wickens, & Hancock, 2015). They all categorize workload as a multi-dimensional construct that is dependent on performance, subjective, and/or physiological responses and “reflect the level of allocation of the specific pool of attentional resources accessed in response to the incipient demands of the task” (Hancock & Matthews, 2019). Research manipulates workload by task complexity, frequency, quantity, and/or allowed completion time. For the purposes of this research, we define workload as the gap between the mental demands placed on the user and his/her attentional resources (Wickens, 2008). We manipulate workload via task quantity due to the long-term goals of UAV command and control (Arrabito et al., 2010; Department of Defense, 2017). Both ends of the workload spectrum have performance challenges: periods of very low workload, i.e., underload, may result in inefficient usage of mental resources (Young & Stanton, 2002) whereas periods of very high workload, i.e., overload, may result in fatigue, frustration, and/or narrowing of attention (Grier et al., 2003; Helton & Russell, 2012).

In contrast, there is no consensus on the associated performance effects for workload transitions, especially when workload transitions from low to high. Evaluating performance during workload transitions often requires separating the workload transition by its periods of low and high workload and then comparing these periods to experimental conditions of constant workload. Alternatively, a subset of studies compares performance before and after a workload transition. These studies are investigating workload history effects, i.e., hysteresis, which examines how a period of previous workload impacts subsequent performance at a different workload level (Goldberg & Stewart, 1981; Farrell, 1999; Matthews, 1986; Morgan & Hancock, 2011). In general, the findings have shown that workload transitions can result in:

- **Performance decrements** (Bowers et al., 2014; Boyer et al., 2015; Cox-Fuenzalida, 2007; Cox-Fuenzalida & Angie, 2005; Cox-Fuenzalida, Angie, et al., 2006; Cox-Fuenzalida, Beeler, et al., 2006; Cumming & Croft, 1973; Goldberg & Stewart, 1980; Hauck et al., 2008; Matthews & Desmond, 2002; Prytz & Scerbo, 2015; Ungar et al., 2005),
- **Performance improvements** (Cox-Fuenzalida, 2007; Cox-Fuenzalida, Swickert, & Hittner, 2004; Cumming & Croft, 1973; Edwards et al., 2017; Goldberg & Stewart, 1980; Hauck et al., 2008; Kim et al., 2019; Krulewitz et al., 1975; Matthews, 1986; Matthews & Desmond, 2002; Prytz & Scerbo, 2015; Ungar et al., 2005),
- **Performance remaining unchanged** (Bowers et al., 2014; Boyer et al., 2015; Helton et al., 2008; Fischer et al., 1995; Matthews, 1986; McKendrick & Harwood, 2019; Prytz & Scerbo, 2015; Voorheis et al., 2005), and
- **Performance fluctuations**, i.e., switching from improving to worsening and vice versa (Bowers et al., 2014; Gluckman et al., 1993; Morgan & Hancock, 2011; Moroney et al., 1995).

Workload transitions are imposed in experiments in several ways. It can consist of changing the event rate of a single task (e.g., Cox-Fuenzalida, 2007), transitioning to and from multiple tasks (e.g., Matthews & Desmond, 2002), or the appearance of an unexpected event (e.g., Boyer et al., 2015). Workload transitions are also studied in very different environments, which include well-validated psychometric tasks (e.g., Bakan vigilance task; Cox-Fuenzalida, 2007) and real-life testbeds (driving simulation; Morgan & Hancock, 201l). However, diverging performance trends exist even within the same experimental setups, suggesting the paradigm is not the primary cause of the divergent performance findings across the literature. For example, performance in a flight simulation appeared highly dependent on the previous period's workload level (Hancock et al., 1995), but studies using the same experimental paradigm failed to replicate this dependency as they found null effects (Fischer et al., 1995; Voorheis et al., 2005). Krulewitz et al. (1975) and Gluckman et al. (1993) also relied on the same workload transition paradigm, but their findings conflicted, which resulted in two different theoretical explanations – a common theme in workload transition research.

To explain the disparate results in single task environments, multiple theoretical explanations have been proposed. These include: expectancy effects (Cumming & Croft, 1973), contrast effects (Krulewitz et al., 1975), short-term memory overload (Goldberg & Stewart, 1980), strategic persistence (Matthews, 1986), motivational intensity theory (Prytz & Scerbo, 2015), and the disruption of cognitive process integration (McKendrick & Harwood, 2019). However, no explanation can apply to all findings (see reviews in Cox-Fuenzalida, 2007 and Gluckman et al., 1993). Studies in multitasking environments mostly rely on two explanations that are rooted in adaptation-based models of mental resources (Hancock & Warm, 1989; Hockey et al., 1986):

1. **Resource depletion.** After a workload transition, there will be a shortage of mental resources and performance will suffer. Resources will eventually recuperate once workload returns to low levels, due to the compensatory regeneration component, and performance will eventually rebound (Gluckman et al., 1993). In other words, there will be a decrement in performance immediately after a shift, but eventually resources will recuperate, and performance will improve.
2. **Effort regulation**. Workload transition performance is dependent on the person accurately appraising, recruiting, and deploying the necessary amount of mental resources for the present workload. Performance is stable as long as their appraisal is correct and workload does not reach levels of overload (Hockey, 1997).

Previous studies support either resource depletion (e.g., Cox-Fuenzalida & Angie, 2005; Gluckman et al., 1993; Moroney et al., 1995), effort regulation (e.g., Cox-Fuenzalida, 2007; Jansen et al., 2016; Matthews & Desmond, 2002), or both depending on operator motivation (Matthews & Desmond, 2002), primary task difficulty (Ungar et al., 2005), or time in experiment (Cox-Fuenzalida, 2007). Clarifying the applicability of theoretical explanations is challenging, which may be why the focus of more recent workload transition research is applied and seldom comments on how the results build upon the existing theory. Bridging the gap between the findings of applied work and their relation to theory not only builds the knowledgebase, but also informs how to design operational environments to properly account for workload transitions.

Current gaps in workload transition research

Currently, applied workload transition research focuses on how features of the environment, transition, and/or individual impact performance. Although diverse, there has been some success in unifying these efforts with human performance models, but their validity and applicability still need extensive research (Sebok et al., 2015). Across these topics, research usually relies on analysis methods based in pairwise comparisons (exceptions: Mracek et al., 2014; McKendrick & Harwood, 2019), which may be contributing to some of the divergence observed in performance trends. The present research addresses these aforementioned factors.

Workload transition research needs to be conducted in applied environments

Studies focused on the characteristics of the environment include examining the effect of alerting participants of an imminent workload transition (Farrell et al., 1999; Helton et al., 2008), modulating primary task difficulty in a dual to single task transition (Matthews & Desmond, 2002; Ungar et al., 2005), providing social support (Hauck et al., 2008), transitioning workload by simulating a total system failure (Morgan & Hancock, 2011), and instructing participants on how to prioritize tasks (Jansen et al., 2016). These studies found performance trends were not affected by external assistance and more dependent on the features of the task itself (Morgan & Hancock, 2011). More recently, studies have examined workload transitions in multitasking environments and the results suggest performance trends can widely vary (Bowers et al., 2013; Cox-Fuenzalida & Angie, 2005; Edwards et al., 2017; Jansen et al., 2016; Kim et al., 2019). Previous studies have shown that high workload performance improves over time and/or performance does not change for any task in the experiment (Bowers et al, 2014; Edwards et al., 2017). Research needs to consider how the completion of multiple tasks may depend on one another or the fluctuating priorities of the environment.

Workload transition research needs to better understand the specifics of the workload transition

Research finds the direction, rate, frequency, and magnitude of a workload transition influences performance effects differently (Cox-Fuenzalida, Beeler, et al., 2006; Matthews, 1986; Morgan & Hancock, 2011; Moroney et al., 1995; Prytz & Scerbo, 2015). Two features of the workload transition that are rarely examined include:

1. **Transition rate**, i.e., the speed at which workload shifts from one load to the next, and
2. **Transition frequency**, i.e., the number of transitions that an individual experiences.

To date, only one study has examined transition rate, as Moroney et al. (1995) compared gradual and abrupt shifts from low to high workload. They found that gradual shifts led to a lower accuracy rates (Moroney et al., 1995). There has also been limited work examining more than two transitions in an experimental scenario (exceptions: Morgan & Hancock, 2011) as most studies transition workload in one (or two) of the paradigms in Figure 1.1.

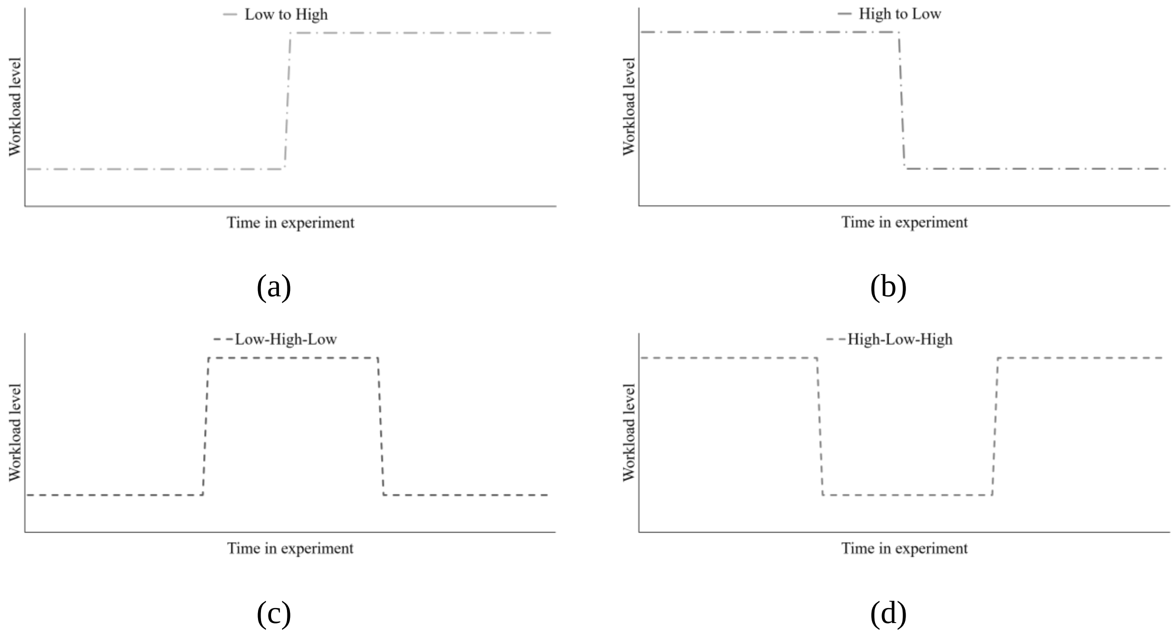


Figure 1.1 Graphical representation of the paradigms often used in workload transition research: (a) an example of a transition from low to high workload, (b) an example of a transition from high to low workload, (c) an example of a transition from low to high back to low workload, and (d) an example of a transition from high to low back to high workload

There are both theoretical and practical motivations to examine the role of transition rate with multiple transitions. First, workload and its transitions are expected to be unpredictable in dynamic domains, so workload transitioning in the same way each time and occurring only once or twice is unlikely. For example, Williams (2006) finds handoffs between UAV operators can increase workload, which is identified as the primary factor for most mishaps, but the way the handoff occurs can vary.

Second, the previous literature has found both theoretical and practical value in studying features of the workload transition, whether it leads to new theoretical explanations (Prytz & Scerbo, 2015) or informs training guidance (Cox-Fuenzalida, Beeler, et al., 2006). For instance, it may be that different transition rates lead to different types of performance trends with regards to speed versus accuracy, low and high workload, and trends over time, as this is observed when manipulating other aspects of the workload transition (e.g., Matthews, 1986).

Workload transition research needs to better account for individual differences

Lastly, the final subset of applied workload transition research focuses on the individual, like the effect of personality (Cox-Fuenzalida, Angie, et al., 2006; Cox-Fuenzalida et al., 2004), working memory capacity (Harwood & McKendrick, 2019), and video game experience (Devlin & Riggs, 2018). In fact, Mracek et al. (2014) successfully predicted performance over time based on the person's perception of how demanding the workload transition was in real-time. Some experimental designs customize the workload transition to an individual's assessed performance capabilities (Bowers et al., 2014; McKendrick & Harwood, 2019), but unexplainable inconsistencies in performance still exist when this method is used. However, the majority of workload transition research still does not account for the individual. This is a practical problem because context-relevant populations, e.g., non-university student populations, are rarely recruited for workload transition research, even though it may matter when studying its performance effects. Edwards et al. (2017) studied workload transition performance of air traffic controllers and performance surprisingly improved, but there was also increased performance variability over time. Unfortunately, their sample size only included eight participants, which makes generalizing their results challenging. Understanding individual differences also has operational value as it can inform the selection and training procedures for UAV command and control (Foroughi et al., 2019).

Workload transition research needs to consider alternative analysis techniques

The divergence in workload transition performance may be due to the method historically used to analyze the data. Most research conducts pairwise comparisons, and although this analysis approach is sufficient to determine workload history effects, it may not be adequate in characterizing how performance depends on time and/or the individual. For example, Jansen et al. (2016) conjectured that they may not have detected a change in performance over time because performance may have been changing *within* each workload period, citing the need for more granular analysis methods. Given performance was averaged by workload period, which is standard in the workload transition literature, there would be no way to detect such a trend. Furthermore, the current analysis methods are unable to measure if the observed performance trends are happening for *all* individuals in the experiment, which could be an alternative explanation to the findings of Jansen et al. (2016) and possibly other studies where discrepancies are observed. Making this distinction is possible with longitudinal data analysis methods, e.g., *growth curve modeling*, which estimates change over time based on how each individual changes over time (Hoffman, 2015).

In summary, there are still several unknowns about workload transitions, which prevent the development of both their theory and design solutions. This dissertation examines how multiple instances of various transition rates affects multitasking performance, specifically in low and high workload, as well as over time. Studying these underexplored features in a realistic setting with novel analyses may fill the current research gaps, while avoiding pitfalls of previous research. At the same time, including measures of the operator that are informative on how the individual manages tasks as workload transitions, may also assist in addressing the current research gaps. For example, including eye tracking data in tandem with, and even as a predictor of, performance aims to provide a richer understanding on the effects of workload transitions.

Motivation to include eye tracking in the present work

Psychophysiological measures have consistently shown the ability to improve the understanding of theoretical explanations, environmental features, and individual differences (see review in Charlton & O'Brien, 2008) – i.e., the present research gaps of workload transitions – but incorporating psychophysiological measures in studies of workload transitions remains limited. A small subset of investigations include cerebral measures and while all reliably track a psychophysiological response during a workload transition, none of the investigations solidly contribute to current theory or design (Bowers et al., 2014; Boyer et al., 2015; Cerruti et al., 2010; Kim et al., 2019; McKendrick & Harwood, 2019). For example, electroencephalograms (EEG) show that the electrical activity in cognitive-related areas of the brain increase as workload increases (such as in Figure 1.1a or 1.1c), suggesting participants actively rely on certain mental resources in the face of workload transitions (Bowers et al., 2014; Kim et al., 2019). However, there were some noted limitations, as EEG is rather invasive, can lead to inconsistent and convoluted interpretations (especially when the performance trends are considered in tandem), and provided no direct guidance on how to best design for the workload transition examined (Bowers et al., 2014; Kim et al., 2019). Hemodynamic measures of the brain have been explored, but it appears they struggle tracking the imposed workload transition (Boyer et al., 2015) even when it is designed specifically for the individual (McKendrick &

Harwood, 2019). Consequently, they added no additional information to the observed performance trends, other than potentially suggesting the activation of the mental resources is not critical to workload transition performance.

Introduction to eye tracking

Eye tracking data may be able to be more specific about the effects of workload transitions as it has been found to provide “objective and quantitative evidence of the user’s visual, overt attentional processes, based on the user’s scan patterns” (Duchowski, 2017, p. 247). This adds another dimension to traditional speed/performance analysis often relied on in human factors research (Duchowski, 2002). Including eye tracking data has been able to directly inform the design of visual displays, given the metrics capture the features of the display most relevant to the attentional process. For example, instead of indicating “cognitive behavior and decision making” were essential to completing the task, eye tracking can specify if visually reading, searching, and/or extracting information was essential and the elements of the display that were related to each action (Duchowski, 2002; Kovesdi et al., 2018; Moacdieh & Sarter, 2017).

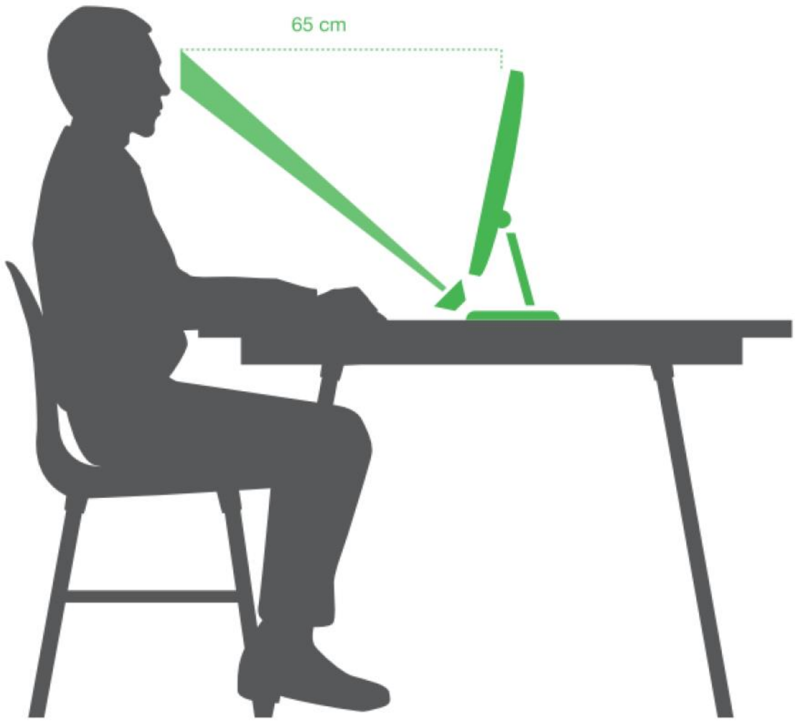
Various human factors research topics have used eye tracking data for theory development, display design, and real-time performance monitoring. Some examples include:

- Information processing (e.g., Duchowski, 2017; Shiferaw et al., 2019),
- Cognitive load (e.g., Coral, 2016; Krejtz et al., 2018; Wilson & Russell, 2007),
- Human-automation trust (e.g., Hergeth et al., 2016; Sarter et al., 2007; Thomas & Wickens, 2004; Victor et al., 2018),
- Individual differences (e.g., Hayes & Henderson, 2017; Jarodzka et al., 2010; Raptis

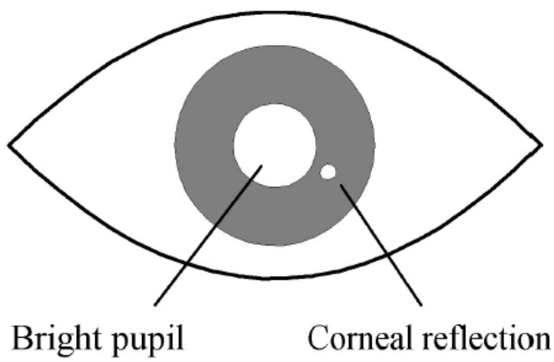
et al., 2017; Shic et al., 2008).

With recent innovations, eye tracking is less invasive, more versatile, more mobile, and more cost-effective than ever before and compared to other psychophysiological measures (e.g., EEG; Dorneich et al., 2008), prompting it for wide-scale used (Krafka et al., 2016; Sibley et al., 2017). Eye tracking is a hopeful tool when pursuing the outstanding research questions in the workload transition literature while also having the potential to be deployed for real-time operational use.

To study scan patterns, eye tracking technology can rely on the corneal-reflection technique (Singh & Sigh, 2012), which shines an infrared light in to a person’s eye to create and track a single reflection point on the cornea (Poole & Ball, 2006; see Figure 1.2 for a schematic).



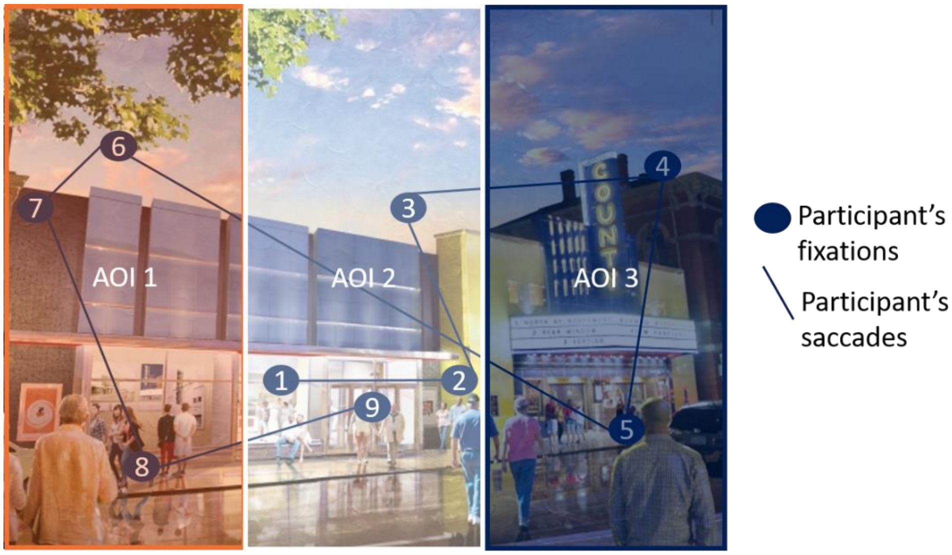
(a)



(b)

Figure 1.2. Experimental setup with the corneal reflection technique: (a) participant seated so eye movements are tracked on the screen (c.f. Gazepoint, 2019) and (b) schematic of the corneal reflection technique (c.f. Poole & Ball, 2006)

The resulting data consists of timestamped Cartesional coordinates, which are based on the eye tracker's sampling rate and the display resolution. These coordinates are then used to distinguish fixations and saccades. Fixations are when the eye is relatively still, allowing for information processing to occur and characterize about 90% of viewing behavior. Saccades are the ballistic movements between fixations (Duchowski, 2007; Poole & Ball, 2006). Fixations and saccades are the basis of most *scan-based* metrics, i.e. measures capturing the features of visual attention allocation (Poole & Ball, 2006). Examples of metrics include the amount of time a given fixation lasts (i.e., fixation duration) or the size of a saccade (i.e., saccade amplitude; Salvucci & Goldberg, 2000). Sometimes, scan-based metrics are calculated and compared across predefined locations on the display, which are termed *areas of interest* (AOIs). Figure 1.3 details how AOIs may be determined on a display, depending on the experimental goals.



(a)

(b)

Figure 1.3. Example of two different AOI discretization methods: (a) context-independent AOIs, i.e., their boundaries were not dependent on features of the image and (b) context-dependent AOIs, i.e., their boundaries depended on semantic features of display (e.g., the theatre, street, and skyline, respectively). It also shows how fixations and saccades may occur across the display, with the numbers indicating the sequential order of each fixation.

Traditionally, eye tracking analyses consist of calculating a set of metrics for the entirety of the experiment and making comparisons based on their average values (Holmqvist et al., 2011, p. 299-468). Recently, there has been interest in creating more advanced scan-based metrics to capture a "higher level descriptor of visual behavior" (Duchowski, 2017, p. 169). For example, discriminating between the individual's reliance on focal (i.e., focused, close together) versus ambient (i.e., global, dispersed) visual attention when completing a visual search task informs the familiarity the individual has with the environment (Irwin & Zelinksy, 2003). Krejtz et al. (2016) created a new scan-based metric based on the normalized size and order of each fixation and saccade as a means to discern focal from ambient visual attention over time. It has since been applied successfully to several different domains and even informs how displays can assist the individual (Krejtz et al., 2017; Lounis et al., 2020). Another set of promising eye tracking metrics are based on the concept of information entropy, as they measure the randomness of visual attention transitions across AOIs and quantitatively explain and compare scan patterns – a constant challenge in the literature (Duchowski, 2017, p. 172; Ellis & Stark, 1986; Shannon, 1948). These metrics have informed concepts including: situation awareness (van de Merwe et al., 2012), decision making (van Maarseveen et al., 2018), task performance (Radun et al., 2017; Shiferaw et al., 2018), task complexity (Di Stasi et al., 2016), and individual differences (Raptis et al., 2017; Shic et al., 2008). This can be particularly informative on task strategy when AOIs are context-dependent (Figure 1.3b).

To the author's knowledge, eye tracking data have never been included in the study of workload transitions. Of immediate interest is to see whether scan-based metrics can help bridge the present research gaps, specifically when it comes to understanding how individuals respond to workload transitions in a multitasking environment (Edwards et al., 2017). Although other measures predict workload transition performance (e.g., perception of the workload transition; Mracek et al., 2014) and non-scan-based metrics successfully measure mental workload (e.g., pupillometry, blink rate; see reviews in Buettner et al., 2018 and Duchowski et al., 2020), these measures do not offer direct information on *how* workload transitions are being managed and *how* the visual display should be designed accordingly. Scan-based metrics have been previously successful in predicting some human factors concepts in applied environments, like situation awareness (Ebeid & Gwizdka, 2018; Ratwani et al., 2010) and working memory capacity (Hayes & Henderson, 2017). If scan-based metrics can predict performance during workload transitions, this would directly and quantifiably address the current research gaps. More generally, it would support the notion that eye tracking may be able to monitor and assist the operator in real-time amongst environment changes, like in the form of an adaptive display (Rothrock et al., 2002; Feigh et al., 2012). Longitudinal analysis methods like growth curve modeling can explore the predictive capability of scan-based metrics directly (Hoffman, 2015).

Motivation and research questions

In order to better detail and design for the effects of workload transitions, it is essential to simultaneously study and connect the mental resources the operator relies on to perform multiple tasks as workload transitions within an applied environment. Specifically, this dissertation explores the role transition rate has on: (a) multitasking performance, (b) performance trends over time, (c) visual attention allocation patterns, and (d) the relation between the aforementioned, in a realistic setting. Although these aspects are prominent in UAV command

and control, very little is known on how they influence the operator's ability to perform in the face of workload transitions. Additionally, addressing each aspect with both traditional and novel measures and analyses aims to innovate both theory and applications of workload transitions. The research questions are as follows:

1. What are the multitasking performance trends of workload transitions and how do they compare to constant workload?
2. What do scan-based eye tracking metrics inform about workload transition performance?
3. How does workload transition rate influence performance trends over time?
4. To what extent are scan-based eye tracking metrics predictive of the performance trends observed over time for workload transitions?

Chapter 2 provides an initial investigation of the multitasking performance when workload transitions at two different rates in a UAV command and control testbed and how it fares to constant workload performance. Chapter 3 expands the understanding of the performance trends observed in Chapter 2 by comparing visual attention allocation patterns across the two transition rates and constant workload by analyzing a suite of scan-based metrics. Chapter 4 is an expansion on the workload transition rate investigation as it studies three different transition rates with a United States Naval aviator population. To accurately detect and characterize how performance trends over time and how it varies across individuals for each transition rate, Chapter 4 analyzes performance with the traditional analysis method, i.e., pairwise comparisons over time and growth curve modeling. Finally, Chapter 5 synthesizes the present work by investigating the predictive ability the scan-based metrics from Chapter 3 have in predicting the performance trends observed in Chapter 4. The goal is to add to the theoretical explanations and provide design guidance surrounding workload transitions for the benefit of multitasking environments.

References

Abich, J., Reinerman-Jones, L., & Matthews, G. (2017). Impact of three task demand factors on simulated unmanned system intelligence, surveillance, and reconnaissance operations. *Ergonomics*, *60*(6), 791–809. https://doi.org/10.1080/00140139.2016.1216171

Arrabito, G. R., Ho, G., Lambert, A., Rutley, M., Keillor, J., Chiu, A., AU, H., & Hou, M. (2010). *Human factors issues for controlling uninhabited aerial vehicles*. Defence Research and Development Canada Toronto. www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA543186

Baker, A. L., & Keebler, J. R. (2017). Factors affecting performance of human-automation teams. In P. Savage-Knepshield & J. Chen (Eds.), *Advances in Human Factors in Robots and Unmanned Systems* (pp. 331–340). https://doi.org/10.1007/978-3-319-41959-6

Bowers, M. A. (2013). *The effects of workload transitions in a multitasking environment*. [Master's thesis, University of Dayton]. OhioLINK.

Bowers, M. A., Christensen, J. C., & Eggemeier, F. T. (2014). The effects of workload transitions in a multitasking environment. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 220–224. https://doi.org/10.1177/1541931214581046

Boyer, M., Cummings, M. L., Spence, L. B., & Solovey, E. T. (2015). Investigating mental workload changes in a long duration supervisory control task. *Interacting with Computers*, *27*(5), 512–520. https://doi.org/10.1093/iwc/iwv012

Buettner, R., Sauer, S., Maier, C., & Eckhardt, A. (2018). Real-time prediction of user performance based on pupillary assessment via Eye Tracking. *AIS Transactions on Human-Computer Interaction*, *10*(1), 26–56. https://doi.org/10.17705/1thci.00102

Cain, B. (2007). *A review of the mental workload literature*. Defence Research and Development

Canada Toronto.

Cerruti, L. M., Estepp, J. R., Miller, W. D., & Christensen, J. C. (2010). Transcranial doppler assessment of workload transition in a complex task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 244–248. https://doi.org/10.1518/107118110X12829369200837

Charlton, S. G., & O'Brien, T. G. (Eds.). (2008). *Handbook of human factors testing and evaluation* (2nd ed.). Lawrence Erlbaum Associates.

Coral, M. P. (2016). *Analyzing cognitive workload through eye-related measurements: A meta-analysis*. [Master's thesis, Wright State University]. CORE Scholar.

Cox-Fuenzalida, L.-E. (2007). Effect of workload history on task performance. *Human Factors*, *49*(2), 277–291. https://doi.org/10.1518/001872007X312496

Cox-Fuenzalida, L.-E., & Angie, A. D. (2005). The effects of workload history on dual task performance. *Current Psychology*, *24*(3), 171–179. https://doi.org/10.1007/s12144-005-1020-y

Cox-Fuenzalida, L.-E., Angie, A., Holloway, S., & Sohl, L. (2006). Extraversion and task performance: A fresh look through the workload history lens. *Journal of Research in Personality*, *40*(4), 432–439. https://doi.org/10.1016/j.jrp.2005.02.003

Cox-Fuenzalida, L.-E., Beeler, C., & Sohl, L. (2006). Workload history effects: A comparison of sudden increases and decreases on performance. *Current Psychology*, *25*(1), 8–14. https://doi.org/10.1007/s12144-006-1012-6

Cox-Fuenzalida, L.-E., Swickert, R., & Hittner, J. B. (2004). Effects of neuroticism and workload history on performance. *Personality and Individual Differences*, *36*(2), 447–456. https://doi.org/10.1016/S0191-8869(03)00108-9

Cummings, M., Bruni, S., Mercier, S., & Mitchell, P. (2007). Automation architecture for single operator, multiple UAV command and control. *The International C2 Journal*, *1*, 1–24. https://doi.org/10.1017/CBO9781107415324.004

Cumming, R. W., & Croft, P. G. (1973). Human information processing under varying task demand. *Ergonomics*, *16*(5), 581–586. https://doi.org/10.1080/00140137308924548

Department of Defense. (2017). *Unmanned systems integrated roadmap FY 2017-2042*.

Devlin, S. P., & Riggs, S. L. (2018). The effect of video game experience and the ability to handle workload and workload transitions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 736–740. https://doi.org/10.1177/1541931218621167

Di Stasi, L. L., Diaz-Piedra, C., Rieiro, H., Sánchez Carrión, J. M., Martin Berrido, M., Olivares, G., & Catena, A. (2016). Gaze entropy reflects surgical task load. *Surgical Endoscopy*, *30*(11), 5034–5043. https://doi.org/10.1007/s00464-016-4851-8

Dorneich, M. C., Mathan, S., Ververs, P. M., & Whitlow, S. D. (2008). Cognitive state estimation in mobile environments. In D. Schmorrow & K. Stanney (Eds.), *Augmented Cognition: A Practitioner's Guide* (pp. 75–111). HFES Press.

Duchowski, A. T. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 455–470. https://doi.org/10.3758/BF03195475

Duchowski, A. T. (2007). *Eye tracking methodology* (2nd ed). Springer-Verlag London.

Duchowski, A. T. (2017). *Eye tracking methodology* (3rd ed). Springer International Publishing AG.

Duchowski, A. T., Krejtz, K., Gehrer, N. A., Bafna, T., & Bækgaard, P. (2020). The low/high index of pupillary activity. *CHI '20*, 1–12. https://doi.org/10.1145/3313831.3376394

Ebeid, I. A., & Gwizdka, J. (2018). Real-time gaze transition entropy. *Proceedings of the 2018*

ACM Symposium on Eye Tracking Research & Applications, 1-3. https://doi.org/10.1145/3204493.3208340

Edwards, T., Gabets, C., Mercer, J., & Biener, N. (2017). Task demand variation in air traffic control: Implications for workload, fatigue, and performance. In: Stanton N., Landry S., Di Bucchianico G., Vallicelli A. (Eds.). *Advances in Human Aspects of Transportation. Advances in Intelligent Systems and Computing*, (pp. 91-102). Springer, Cham. https://doi.org/10.1007/978-3-319-41682-3_8

Edwards, T., Sharples, S., Wilson, J. R., & Kirwan, B. (2012). Factor interaction influences on human performance in air traffic control: The need for a multifactorial model. *Work*, *41*, 159–166. https://doi.org/10.3233/WOR-2012-0151-159

Ellis, S. R., & Stark, L. (1986). Statistical dependency in visual scanning. *Human Factors*, *28*(4), 421–438. https://doi.org/10.1177/001872088602800405

Farrell, P. S. E. (1999). The hysteresis effect. *Human Factors*, *41*(2), 226–240. https://doi.org/10.1518/001872099779591259

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. https://doi.org/10.1177/0018720812443983

Fischer, D. S. (1995). *The effect of previous levels of workload in a simulated flight task.* [Unpublished master's thesis]. University of Dayton.

Foroughi, C. K., Rovira, E., Rose, K., Davis, D., Jurado, J. J., Chae, J., Pak, R., Brown, N. L., Sibley, C., & Coyne, J. T. (2019). How do individual differences in attention control and working memory capacity relate to monitoring automation? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 1409–1413.

https://doi.org/10.1177/1071181319631196

Freedberg Jr., S. J. (2021, June 7). Drone refuels manned aircraft for first time: Navy. *Breaking Defense*. https://breakingdefense.com/2021/06/drone-refuels-manned-aircraft-for-first-time-navy/

Gaillard, A. W. K., & Kramer, A. F. (2009). Theoretical and methodological issues in psychophysiological research. In R. W. Backs & W. Boucsein (Eds.), *Engineering psychophysiology: Issues and applications*. Lawrence Erlbaum Associates, Inc.

Gazepoint Control user manual. (2019). https://www.gazept.com/

Giese, S., Carr, D., & Chahl, J. (2013). Implications for unmanned systems research of military UAV mishap statistics. *IEEE Intelligent Vehicles Symposium (IV)*, 1191–1196. https://doi.org/10.1109/IVS.2013.6629628

Gluckman, J. P., Warm, J. S., Dember, W. N., & Rosa, R. R. (1993). Demand transitions and sustained attention. *The Journal of General Psychology*, *120*(3), 323–337. https://doi.org/10.1080/00221309.1993.9711151

Goldberg, R. A., & Stewart, M. R. (1980). Memory overload or expectancy effect? ‘Hysteresis’ revisited*. *Ergonomics*, *23*(12), 1173–1178. https://doi.org/10.1080/00140138008924824

Goodrich, M. A., & Cummings, M. L. (2015). A human factors perspective on next generation unmanned aerial systems. In K. Valavanis & G. Vachtsevanos (Eds.), *Handbook of Unmanned Aerial Vehicles* (pp. 2405–2423). Springer, Dordrecht.

Grier, R. A., Warm, J. S., Dember, W. N., Matthews, G., Galinsky, T. L., Szalma, J. L., & Parasuraman, R. (2003). The vigilance decrement reflects limitations in effortful attention, not mindlessness. *Human Factors*, *45*(3), 349–359. https://doi.org/10.1518/hfes.45.3.349.27253

Grier, R. A., Wickens, C. D., Kaber, D., Strayer, D., Boehm-Davis, D., Trafton, J. G., St. John, M. (2008). The red-line of workload: Theory, research, and design. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1204–1208. https://doi.org/10.1177/154193120805201811

Hancock, P. A., & Matthews, G. (2019). Workload and performance: Associations, insensitivities, and dissociations. *Human Factors*, *61*(3), 374–392. https://doi.org/10.1177/0018720818809590

Hancock, P. A., & Warm, J. S. (1989). A dynamic model of stress and sustained attention. *Human Factors*, *31*(5), 519–537. https://doi.org/10.7771/2327-2937.1024

Hancock, P. A., Williams, G., Manning, C. M., & Miyake, S. (1995). Influence of task demand characteristics on workload and performance. *The International Journal of Aviation Psychology*, *5*(1), 63–86. https://doi.org/10.1207/s15327108ijap0501_5

Hauck, E. L., Snyder, L. A., & Cox-Fuenzalida, L. E. (2008). Workload variability and social support: Effects on stress and performance. *Current Psychology*, *27*(2), 112–125. https://doi.org/10.1007/s12144-008-9026-x

Hayes, T. R., & Henderson, J. M. (2017). Scan patterns during real-world scene viewing predict individual differences in cognitive capacity. *Journal of Vision*, *17*(5). https://doi.org/10.1167/17.5.23

Helton, W. S., & Russell, P. N. (2012). Brief mental breaks and content-free cues may not keep you focused. *Experimental Brain Research*, *219*(1), 37–46. https://doi.org/10.1007/s00221-012-3065-0

Helton, W. S., Shaw, T., Warm, J. S., Matthews, G., & Hancock, P. (2008). Effects of warned and unwarned demand transitions on vigilance performance and stress. *Anxiety, Stress, and*

Coping, *21*(2), 173–184. https://doi.org/10.1080/10615800801911305

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human Factors*, *58*(3), 509–519. https://doi.org/10.1177/0018720815625744

Hobbs, A., & Shively, R. J. (2014). Human factor challenges of remotely piloted aircraft. *31st EAAP Conference*, 5–14.

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive energetical framework. *Biological Psychology*, *45*, 73–93. https://doi.org/10.1016/S0301-0511(96)05223-4

Hockey, G. R. J., Coles, M. G. H., & Gaillard, A. W. K. (1986). Energetical issues in research on human information processing. In G. R. J. Hockey, A. W. K. Gaillard, & M. G. H. Coles (Eds.), *Energetics and Human Information Processing*. NATO ASI Series (Series D: Behavioural and Social Sciences). Springer, Dordrecht. https://doi.org/10.1007/978-94-009-4448-0_1

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. Routledge.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods, paradigms, and measures*. Oxford University Press.

Huey, B. M., Wickens, C. D. (Eds.). (1993). *Workload transition: Implications for individual and team performance*. National Research Council.

Irwin, D. E., & Zelinsky, G. J. (2002). Eye movements and scene perception: Memory for things observed. *Perception and Psychophysics, 64*(6), 882–895.

https://doi.org/10.3758/BF03196793

Jansen, R. J., Sawyer, B. D., van Egmond, R., de Ridder, H., & Hancock, P. A. (2016). Hysteresis in mental workload and task performance: The influence of demand transitions and task prioritization. *Human Factors*, *58*(8), 1143–1157. https://doi.org/10.1177/0018720816669271

Jarodzka, H., Scheiter, K., Gerjets, P., & van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learning and Instruction*, *20*(2), 146–154. https://doi.org/10.1016/j.learninstruc.2009.02.019

Kim, N., House, R., Yun, M. H., & Nam, C. S. (2019). Neural correlates of workload transition in Multitasking: An ACT-R model of hysteresis effect. *Frontiers in Human Neuroscience*, *12*(535), 1–14. https://doi.org/10.3389/fnhum.2018.00535

Kovesdi, C., Spielman, Z., Leblanc, K., & Rice, B. (2018). Application of eye tracking for measurement and evaluation in human factors studies in control room modernization. *Nuclear Technology*, *202*(2–3), 220–229. https://doi.org/10.1080/00295450.2018.1455461

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., & Torralba, A. (2016). Eye tracking for everyone. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2176-2184. http://gazecapture.csail.mit.edu

Krejtz, K., Çöltekin, A., Duchowski, A. T., & Niedzielska, A. (2017). Using coefficient *K* to distinguish ambient/focal visual attention during cartographic tasks. *Journal of Eye Movement Research*, *10*(2). https://doi.org/10.16910/jemr.10.2.3

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), 1–23. https://doi.org/10.1371/journal.pone.0203629

Krejtz, K., Duchowski, A., Krejtz, I., Szarkowska, A., & Kopacz, A. (2016). Discerning ambient/focal attention with coefficient *K*. *ACM Transactions on Applied Perception*, *13*(3), 1–20. https://doi.org/10.1145/2896452

Krulewitz, J. E., Warm, J. S., & Wohl, T. H. (1975). Effects of shifts in the rate of repetitive stimulation on sustained attention. *Perception & Psychophysics*, *18*(4), 245–249. https://doi.org/10.3758/BF03199369

Lounis, C. A., Hassoumi, A., Lefrancois, O., Peysakhovich, V., & Causse, M. (2020). Detecting ambient/focal visual attention in professional airline pilots with a modified coefficient *K*: A full flight simulator study. *Eye Tracking Research & Applications Symposium (ETRA)*. https://doi.org/10.1145/3379157.3391412

Matthews, M. L. (1986). The influence of visual workload history on visual performance. *Human Factors*, *28*(6), 623–632. https://doi.org/10.1177%2F001872088602800601

Matthews, G., & Desmond, P. A. (2002). Task-induced fatigue states and simulated driving performance. *The Quarterly Journal of Experimental Psychology*, *55A*(2), 659–686. https://doi.org/10.1080/0272498014300050

McCarley, J. S., & Wickens, C. D. (2004). *Human factors concerns in UAV flight*. Technical Report AHFD-05-05/FAA-05-01. Aviation Human Factors Division Institute of Aviation, University of Illinois. Savoy, Illinois, USA.

McKendrick, R., & Harwood, A. (2019). Cognitive workload and workload transitions elicit curvilinear hemodynamics during spatial working memory. *Frontiers in Human Neuroscience*, *13*(November), 1–16. https://doi.org/10.3389/fnhum.2019.00405

Moacdieh, N. M., & Sarter, N. (2017). The effects of data density, display organization, and stress on search performance: An eye tracking study of clutter. *IEEE Transactions on*

Human-Machine Systems, *47*(6), 886–895. https://doi.org/10.1109/THMS.2017.2717899

Moon, J., Choe, S., Lee, S., & Kwon, O. S. (2019). Temporal dynamics of visual attention allocation. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-40281-7

Morgan, J. F., & Hancock, P. A. (2011). The effect of prior task loading on mental workload: An example of hysteresis in driving. *Human Factors*, *53*(1), 75–86. https://doi.org/10.1177/0018720810393505

Moroney, B. W., Warm, J. S., & Dember, W. N. (1995). Effects of demand transitions on vigilance performance and perceived workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1375–1379. https://doi.org/10.1177%2F154193129503902104

Mracek, D. L., Arsenault, M. L., Day, E. A., Hardy, J. H., & Terry, R. A. (2014). A multilevel approach to relating subjective workload to performance after shifts in task demand. *Human Factors*, *56*(8), 1401–1413. https://doi.org/10.1177/0018720814533964

Murray, C. A., & Cox, C. B. (1989). *Apollo: The race to the moon*. Simon & Schuster.

Parush, A., & Rustanjaja, N. (2013). The impact of sudden events and spatial configuration on the benefits of prior information to situation awareness and performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1395–1399. https://doi.org/10.1177/1541931213571311

Poole, A., & Ball, L. J. (2006). Eye tracking in human-computer interaction and usability research: Current status and future prospects. *Encyclopedia of Human-Computer Interaction*, 211–219. https://doi.org/10.4018/978-1-59140-562-7.ch034

Prytz, E. G., & Scerbo, M. W. (2015). Changes in stress and subjective workload over time following a workload transition. *Theoretical Issues in Ergonomics Science*, *16*(6), 586–605.

https://doi.org/10.1080/1463922X.2015.1084397

Radun, J., Nuutinen, M., Antons, J. N., & Arndt, S. (2017). Did you notice it-how can we predict the subjective detection of video quality changes from eye movements? *IEEE Journal on Selected Topics in Signal Processing*, *11*(1), 37–47. https://doi.org/10.1109/JSTSP.2016.2607696

Raptis, G., Katsini, C., Belk, M., Fidas, C., Samaras, G., & Avouris, N. (2017). Using eye gaze data and visual activities to infer human cognitive styles: Method and feasibility studies. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, 164–173. https://doi.org/10.1145/3079628.3079690

Ratwani, R. M., McCurry, J. M., & Trafton, J. G. (2010). Single operator, multiple robots: An eye movement based theoretic model of operator situation awareness. *HRI '10 Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, *55*, 235–242. https://doi.org/10.1109/HRI.2010.5453191

Rothrock, L., Koubek, R., Fuchs, F., & Haas, M. (2002). Reappraisal of adaptive interfaces: Toward biologically inspired paradigms. *Theoretical Issues in Ergonomics Science*, *3*(May), 47–84. https://doi.org/10.1080/14639220110110342

Sarter, N. B., Mumaw, R. J., & Wickens, C. D. (2007). Pilots' monitoring strategies and performance on automated flight decks: An empirical study combining behavioral and eye-tracking data. *Human Factors*, *49*(3), 347–357. https://doi.org/10.1518/001872007X196685

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the Eye Tracking Research & Applications Symposium*, 71–78. https://doi.org/10.1145/355017.355028

Sebok, A., Wickens, C. D., & Sargent, R. (2015). Development, testing, and validation of a model-based tool to predict operator responses in unexpected workload transitions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 622–626. https://doi.org/10.1177/1541931215591136

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Shic, F., Chawarska, K., Bradshaw, J., & Scassellati, B. (2008). Autism, eye-tracking, entropy. *7th IEEE International Conference on Development and Learning, ICDL*, 73–78. https://doi.org/10.1109/DEVLRN.2008.4640808

Shiferaw, B., Downey, L., & Crewther, D. (2019). A review of gaze entropy as a measure of visual scanning efficiency. *Neuroscience and Biobehavioral Reviews*, *96*, 353–366. https://doi.org/10.1016/j.neubiorev.2018.12.007

Shiferaw, B., Downey, L. A., Westlake, J., Stevens, B., Rajaratnam, S. M. W., Berlowitz, D. J., Swann, P., & Howard, M. E. (2018). Stationary gaze entropy predicts lane departure events in sleep-deprived drivers. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-20588-7

Sibley, C., Coyne, J., & Morrison, J. (2015). Research considerations for managing future unmanned systems. *AAAI Spring Symposium*, 54–59.

Sibley, C. M., Foroughi, C. K., Olson, T., Moclaire, C., & Coyne, J. T. (2017). Practical considerations for low-cost eye tracking: An analysis of data loss and presentation of a solution. In D. Schmorrow & C. Fidopiastis (Eds.), *International Conference on Augmented Cognition*, (pp. 236–250). Springer, Cham. https://doi.org/10.1007/978-3-319-58628-1

Singh, H., & Singh, J. (2012). Human eye tracking and related issues: A review. *International Journal of Scientific and Research Publications*, *2*(9).

Thomas, L. C., & Wickens, C. D. (2004). Eye-tracking and individual differences in off-normal event detection when flying with a synthetic vision system display. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 223–227. https://doi.org/10.1177/154193120404800148

Ungar, N. R., Matthews, G., Warm, J. S., Dember, W. N., Thomas, J. K., Finomore, V. S., & Shaw, T. H. (2005). Demand transitions and tracking performance efficiency: Structural and strategic models. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1523–1526. https://doi.org/10.1177/154193120504901704

Van Acker, B. B., Parmentier, D. D., Vlerick, P., & Saldien, J. (2018). Understanding mental workload: From a clarifying concept analysis toward an implementable framework. *Cognition, Technology and Work*, *20*(3), 351–365. https://doi.org/10.1007/s10111-018-0481-3

van Maarseveen, M. J. J., Oudejans, R. R. D., Mann, D. L., & Savelsbergh, G. J. P. (2018). Perceptual-cognitive skill and the in situ performance of soccer players. *Quarterly Journal of Experimental Psychology*, *71*(2), 455–470. https://doi.org/10.1080/17470218.2016.1255236

van de Merwe, K., van Dijk, H., & Zon, R. (2012). Eye movements as an indicator of situation awareness in a flight simulator experiment. *International Journal of Aviation Psychology*, *22*(1), 78–95. https://doi.org/10.1080/10508414.2012.635129

Victor, T. W., Tivesten, E., Gustavsson, P., Johansson, J., Sangberg, F., & Ljung Aust, M. (2018). Automation expectation mismatch: Incorrect prediction despite eyes on threat and hands on wheel. *Human Factors*, *60*(8), 1095–1116. https://doi.org/10.1177/0018720818788164

Voorheis, C. M., Moroney, W. F., Biers, D. W., & Eggemeier, F. T. (2005). In search of workload context effects using a compensatory tracking task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2095–2098. https://doi.org/10.1177/154193120504902406

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*(3), 449–455. https://doi.org/10.1518/001872008X288394

Williams, K. W. (2006). *Human factors implications of unmanned aircraft accidents: Flight-control problems.* FAA Civil Aerospace Medical Institute. https://doi.org/10.1016/S1479-3601(05)07008-6

Wilson, G. F., & Russell, C. A. (2007). Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. *Human Factors*, *49*(6), 1005–1018. https://doi.org/10.1518/001872007X249875.1006

Woods, D. D., Patterson, E. S., & Roth, E. M. (2002). Can we ever escape from data overload? A cognitive systems diagnosis. *Cognition, Technology, & Work*, *4*(1), 22–36.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17. https://doi.org/10.1080/00140139.2014.956151

Young, M. S., & Stanton, N. A. (2002). Malleable attentional resources theory: A new explanation for the effects of mental underload on performance. *Human Factors*, *44*(3), 365–375. https://doi.org/10.1518/0018720024497709

CHAPTER 2

What are the multitasking performance trends of workload transitions and how do they compare to constant workload?

Introduction

As Chapter 1 details, complex and dynamic environments require operators to transition seamlessly between varying levels of workload. For example, operators overseeing the command and control of unmanned aerial vehicles (UAVs) are subject to varying workload levels as they manage various responsibilities, task demands, and automation levels (Sibley et al., 2015). To date, most research has focused on the effects of low or high workload; however, transitions from low to high workload are not studied nearly as often, even though they are more realistic to what operators experience while on the job, and are increasingly probable due to the increased reliance on automation (Baker & Keebler, 2017; Hooey et al., 2017; Huey & Wickens, 1993). In this chapter, we investigated the effects of medium and fast transitions from low to high workload in a dynamic, multitasking environment, to better understand the effect these transitions have on performance in complex and dynamic domains. We also compared performance of the two transition rates over time, to understand the presence and severity of workload history effects.

As discussed in Chapter 1, there is currently no consensus on the associated performance effects for workload transitions. Subsequently, theoretical explanations of workload history are not widely agreed upon (Cox-Fuenzalida, 2007). Given that the impact of workload transitions may be dynamic, further examination of transition effects on performance is needed (Jansen et al., 2016). For example, performance has been found to be dependent on the direction of the transition (i.e., transitions from low to high versus high to low workload; Cumming & Croft, 1973; Cox-Fuenzalida, Beeler, et al., 2006), the magnitude of the transition (Prytz & Scerbo, 2015), and measures of performance (errors of commission versus omission; Cox-Fuenzalida, 2007). Less is known on how the rate of a workload transition affects multitasking performance. To our knowledge, only one study to date has compared gradual and sudden shifts from low to high workload, and found that gradual shifts led to a lower accuracy rates (Moroney et al., 1995). More work is needed to further understand these performance trends, especially for domains requiring multitasking.

Several theoretical explanations have been proposed for performance where workload transitions in single-task environments (see review in Cox-Fuenzalida, 2007; McKendrick & Harwood, 2019; Prytz & Scerbo, 2015). Two explanations have been primarily used to explain performance when transitioning between a dual- and single-task paradigm: resource depletion and effort regulation (see Chapter 1 for definitions). Previous studies support either resource depletion (Cox-Fuenzalida & Angie, 2005; Gluckman et al., 1993; Moroney et al., 1995), effort regulation (Cox-Fuenzalida, 2007; Jansen et al., 2016; Matthews & Desmond, 2002), or both (Matthews & Desmond, 2002; Ungar et al., 2005). Across these studies, performance during workload transitions is either compared to constant workload (e.g., Bowers et al., 2014; Cox-Fuenzalida, 2007; Moroney et al., 1995; Ungar et al., 2005) or over time (e.g., Jansen et al., 2016; Kim et al., 2019; Morgan & Hancock, 2011). Rarely are both analyses conducted, which may be contributing to the divergence across the literature. For example, performance trends may not be the same for both low and high workload as some previous work shows their performance trends match (Cox-Fuenzalida, 2007; Cox-Fuenzalida, Angie, et al., 2006; Cox-Fuenzalida & Angie, 2005; Helton et al., 2008; Moroney et al., 1995), but others have found it differs (e.g., decrements for low workload, but not high workload; Cox-Fuenzalida, 2007; Matthews, 1986; Prytz & Scerbo, 2015). Also, performance differences may not be immediate as an operator may be able to manage a workload transition, but it may have ramifications on performance at a later point in time (Edwards et al., 2017; Huey & Wickens, 1993).

This chapter aims to address some of the knowledge gaps in workload transition research, particularly as it relates to complex environments:

1. **Transition rate:** The speed at which workload transitions has been underexplored. Most research has focused on immediate shifts between low and high workload, whereas there has been less reported on shifts that are not instant (exception: Matthews, 1986; Moroney et al., 1995). When it has been explored, performance has only been compared to constant workload and not to other transitions rates nor over time.
2. **Theoretical explanations:** There is also a need to understand the applicability and validity of the resource depletion and effort regulation explanations in multitasking domains that are not just transitioning between dual- and single-task paradigms (Cox-Fuenzalida, 2007). Existing evidence supports both explanations (Ungar et al., 2005), but most investigations have been limited to studying a single dual- to single-task transition. Previous work has explored the applicability of these explanations in other types of multitasking environments (Bowers et al., 2014; Jansen et al., 2016), but none test their applicability directly.
3. **Realistic environment features**: Historically, factors pertinent to complex and

dynamic work environments are underexamined in the workload transition literature (Cox-Fuenzalida & Angie, 2005; Huey & Wickens, 1993). This includes better understanding multitasking performance and the impact of multiple transitions.

a. Previous researchers have suggested that findings from single-task studies may not hold true in multitasking environments (Gluckman et al., 1993), so these environments warrant their own investigation. When multitasking has been included in workload transition studies, it has been in either highly-controlled laboratory settings (Cox-Fuenzalida & Angie, 2005), realistic testbeds (Jansen et al., 2016; Morgan & Hancock, 2011), or thoroughly validated multitasking environments (e.g., Air Force Multi-Attribute Task Battery; Bowers et al., 2014). Of specific interest is to further explore primary and secondary task performance effects during workload transitions, (Cox-Fuenzalida & Angie, 2005; Jansen et al., 2016; Morgan & Hancock, 2011), especially when they are realistically related. For example, operators in these environments are expected to continuously manage various interdependent tasks and responsibilities that fluctuate in their demand for operator attention. Depending on the situation, tasks that may be secondary to the mission may be critical to the overall viability of the UAV, so capturing this nuance in task prioritization is important (Clare et al., 2010). Understanding the effects of secondary tasks is important to not only expand the current workload transition literature, but also as operational environments become increasingly complex and greater threaten operator and system safety (Cummings, 2014; Matthews, et al., 1996).

b. Also, there is a need to understand how multiple workload transitions impact performance trends. The majority of existing research has employed one or two workload transitions in a single-task environment (an exception is the work by Morgan & Hancock, 2011); however, findings from one or two workload transitions may not be representative of the multiple transitions operators may experience while on the job.

The goal of Chapter 2 is to address these research gaps simultaneously. The three specific research questions (RQ) are:

1. **RQ 2.1:** How does performance compare between medium and fast transitions from low to high workload?
   a. **Expectations:** In order to answer this question, we will compare performance between two transition rates as a whole, per workload level, and per workload period, i.e., study its workload history effect. We expect fast transitions will result in better primary and secondary task performance compared to medium transitions (Moroney et al., 1995). As for how it will compare over time, we expect primary task performance to degrade with workload transitions (Cox-Fuenzalida, 2007; Cox-Fuenzalida & Angie, 2005; Hancock et al., 1995; Morgan & Hancock, 2011) and it to be more pronounced for medium transitions compared to fast ones (Moroney et al., 1995). Performance may partially recover and then plateau across the low workload periods, as this is consistent with the previous research that studies workload transitions in dynamic, multitasking environments (Edwards et al., 2017; Jansen et al., 2016; Morgan & Hancock, 2011).

2. **RQ 2.2:** How does performance compare between the low workload periods of medium and fast transitions and constantly low workload?
    a. **Expectations:** Previous work has shown that both primary and secondary task performance were better when workload was constantly low (Bowers et al., 2014; Cox-Fuenzalida, 2007; Cox-Fuenzalida & Angie, 2005; Matthews, 1986; Ungar et al., 2005; Wickens et al., 1985). Additionally, the resource depletion explanation predicts that both primary and secondary task performance will be better when workload is constantly low, compared to the low workload periods of medium and fast transitions, because resources will have not been depleted from any transitions during constant low workload.
3. **RQ 2.3:** How does performance compare between the high workload periods of medium and fast transitions and constantly high workload?
    a. **Expectations:** We expected that performance will be worse for both primary and secondary tasks when workload is constantly high (Bowers et al., 2014; Cox-Fuenzalida & Angie, 2005; Matthews, 1986; Prytz & Scerbo, 2015; Wickens et al., 1985). The effort regulation explanation predicts that primary task performance will be better during the high workload periods of medium and fast transitions than constant high workload, because the workload transitions will signal for the adjustment of resources accordingly.

Table 2.1 summarizes how each research question expands on each theoretical explanation. Of note, RQ 2.1 is not included because it does not provide further insights into the theoretical explanations.

Table 2.1 Justification of how research questions to map to theoretical explanations

Research Question (RQ)	Resource Depletion Explanation	Effort Regulation Explanation
RQ 2.2: Low workload vs. the low workload periods of medium and fast transitions, respectively	<u>If constant low workload outperforms the low workload periods of medium or fast transitions:</u> • Resources depletion <u>**IS**</u> a function of workload transitions. Workload transitions deplete more mental resources than constant low workload (Bowers et al., 2014; Cox-Fuenzalida, 2007; Cox-Fuenzalida & Angie, 2005; Matthews, 1986; Ungar, 2005). <u>If the low workload periods of medium or fast transitions is equal to or outperforms low workload:</u> • Resources depletion <u>**IS NOT**</u> a function of workload transitions. Workload transitions do not deplete mental resources differently than low workload (Matthews, 1986).	n/a
RQ 2.3: Constant high workload vs. the high workload periods of medium and fast workload transitions, respectively	n/a	<u>If constant high workload is equal to or outperforms the high workload period of medium and fast transitions:</u> • Effort regulation <u>**IS NOT**</u> a function of workload transitions. Workload transitions do not impact the effective regulation of mental resources (Gluckman et al., 1993). <u>If the high workload period of medium and fast transitions outperforms constant high workload:</u> • Effort regulation <u>**IS**</u> a function of workload transition. Workload transitions can help actively regulate mental resources under high workload (Hockey, 1997; Matthews & Desmond, 2002; Matthews, 1986).

The goal of this chapter is to gain an initial understanding on how different workload transition rates impacts both primary and secondary task performance, and how that compares to constant workload and over time. It's the first step in informing how to design a display to better account for workload transition effects, as a means to make complex domains safer and more effective. Unmanned aerial vehicle (UAV) command and control was specifically of interest per the current and expected demands of the environments and the initiative to address these challenges with design solutions (Department of Defense, 2017).

Method

Participants

Twenty-one students participated in this study (13 males and 8 females; *M*=20.9 years, *SD*=1.5 years) and each was compensated $10/hour. The data of two participants was excluded from the workload history analysis: one had missing eye tracking data (which will be important for the analysis in Ch. 3) and another had no primary task performance data during one of the low workload periods. The study was approved by the Clemson University Institutional Review Board (IRB2015-217) and all participants provided informed consent.

Experimental setup

The testbed was developed using the Unity game development platform and was based on the 'Vigilant Spirit Control Station' (VSCS) used by the Air Force to develop interfaces to control multiple UAVs (Feitshans et al., 2008). The VSCS platform has been used to test how interface design could aid UAV operators. Tasks in the testbed were typical of an UAV command and control environment, such as target detection and route planning (Feitshans et al., 2008), which require operators to employ perceptual, cognitive, and motor resources. The testbed ran on a desktop computer with a Dell 32” monitor (2560×1600 resolution) and a standard mouse.

UAV command and control testbed

Participants were responsible for controlling and managing UAVs under four 15-minute scenarios in the UAV testbed, i.e., testbed scenario. There were four tasks in each scenario, one primary task and three secondary tasks, and each will be discussed in turn, but Figure 2.1 shows the interface of the testbed. For each scenario, the frequency of the primary task varied (see section **Testbed scenarios**), while one of the three secondary tasks occurred every 20 s, on average, in a pseudo-random order.

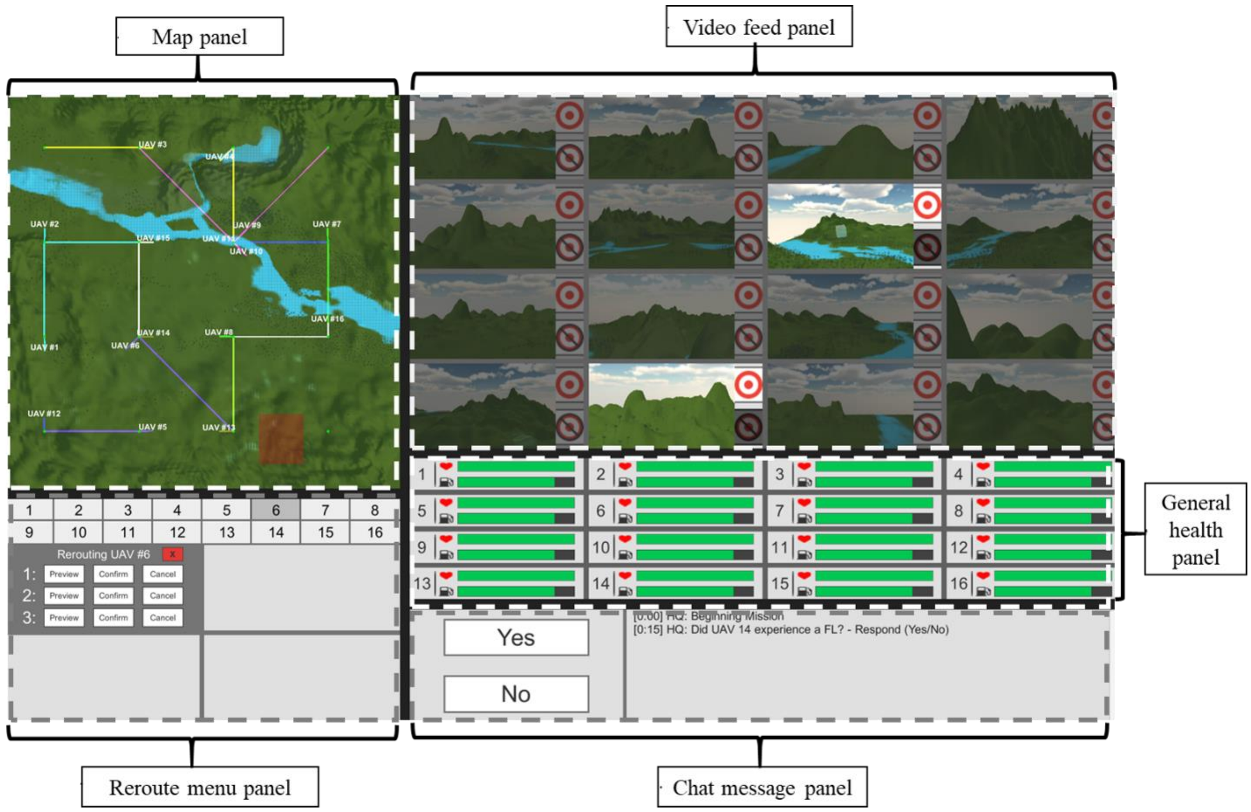


Figure 2.1 The interface of the testbed used in the study

Target detection task (primary task)

Participants were instructed that this task had the highest priority. They were tasked to monitor up to 16 UAV video feeds on the Video Feed panel for a target, presented as a semi-transparent cube (see Figure 2.2). Targets could only be detected when a UAV video feed was active (illuminated). When a UAV video feed was active and a target was present, participants were instructed to press the “target” button. Otherwise, they were instructed to leave the default “no target” button selected. The “no target” button was the default, as pilot testing suggested this better assessed the participant’s search abilities and not their ability to click quickly. UAV video feeds were active for 10 s, and targets could appear during this time. Video feeds cycled between active and inactive throughout the scenario. If a target was present in an active UAV, but the participant did not select the target button within the 10 s, the participant missed the opportunity to detect that specific target. On average, 20% of active UAVs detected a target. The number of simultaneously active UAVs determined the workload level (see section **Testbed scenarios**).

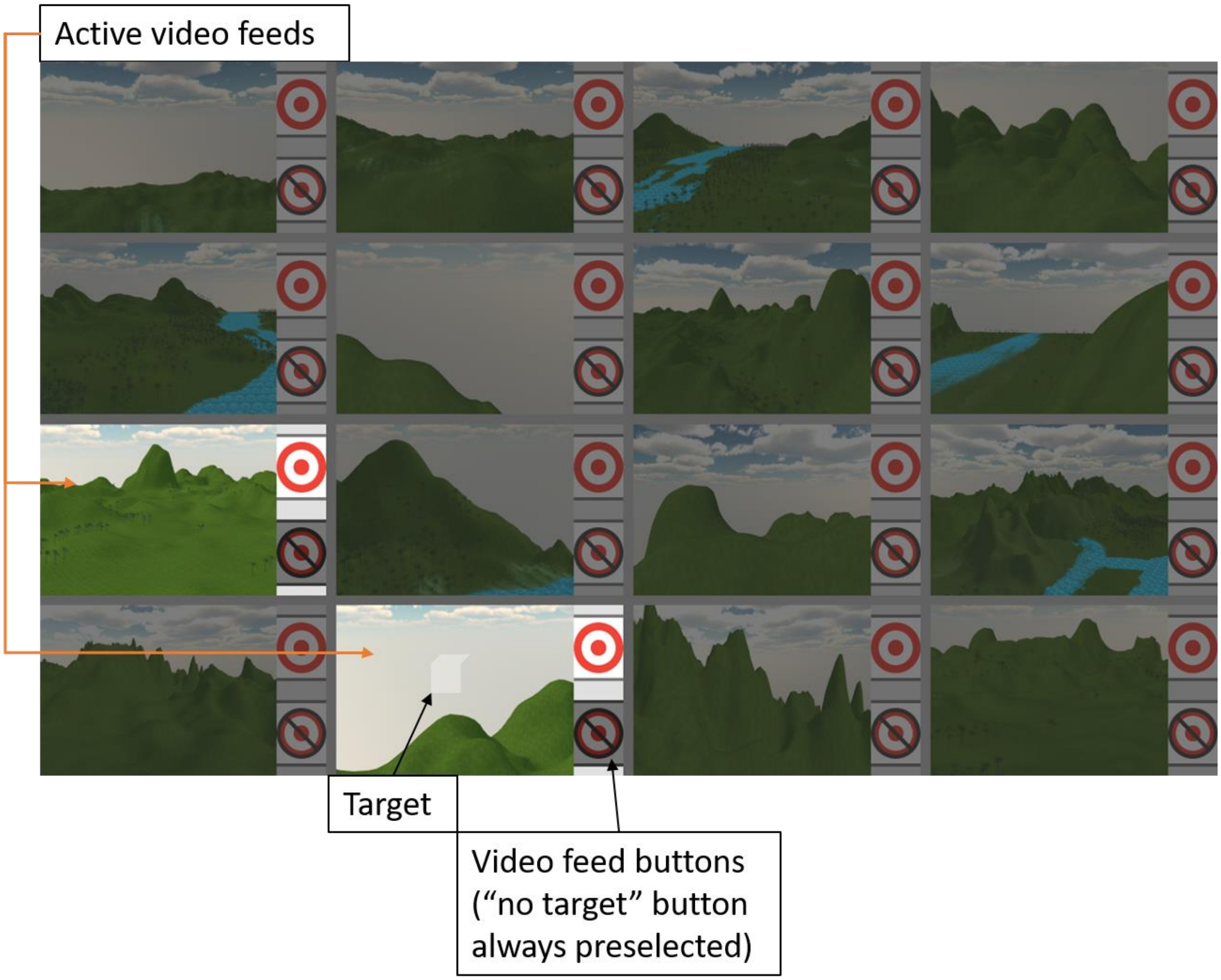


Figure 2.2 Example of active and inactive UAVs on the Video Feed panel and directions on how to detect a target

Reroute task (secondary task)

Participants were tasked to reroute a UAV when it was projected to enter a no-fly-zone (i.e., red square on the Map panel in Figure 2.3). If a UAV was projected to enter the no-fly-zone, its route and label would turn orange and participants had 15–20 s to reroute it away from the no-fly-zone. To reroute a UAV, a participant clicked on the respective UAV's numbered square in the Reroute Menu panel and chose from one of three new routes. For each new route, participants could select "Preview" to see the alternative route, "Confirm" to reroute the UAV to a specific alternative route, or "Cancel" to exit from previewing the alternative route. There was no limit to how many times a UAV could be rerouted. If a UAV was not rerouted to avoid the no-fly-zone, it became nonoperational for the remainder of the scenario. The rerouting task occurred 18 times in each scenario.

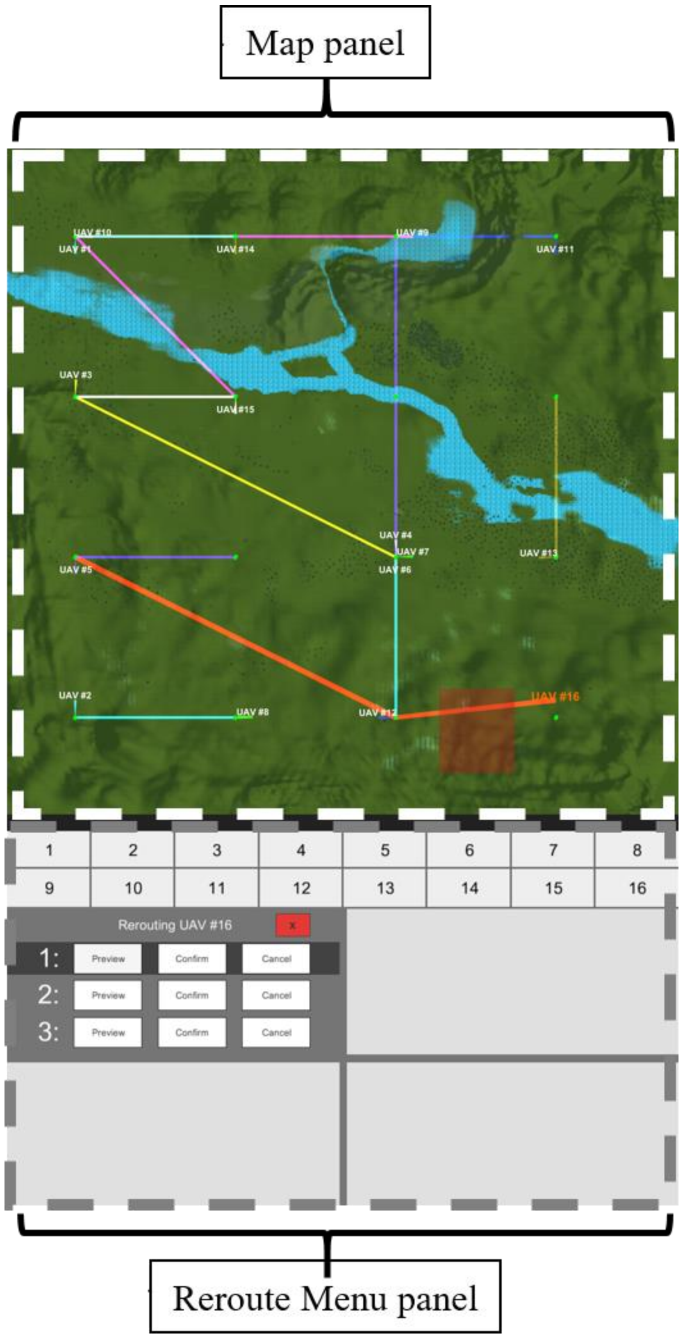


Figure 2.3 Map panel (top half) shows projected routes for each UAV and Reroute Menu panel (bottom half) is where participants could choose a new route for a given UAV. After clicking on the UAV's number from the buttons in the top two rows in Reroute Menu panel (i.e., the buttons numbered 1–16), a menu of route options was presented. The "Preview" button allowed participants to see if the alternative route, which was overlaid on the Map panel, avoided the no-

fly-zone, the “Confirm” button reroute the UAV to that alternative route, and the “Cancel” button removed the overlaid alternative route from the Map panel

Fuel leak task (secondary task)

Participants were also tasked to monitor for fuel leaks using the General Health panel (Figure 2.4). When a fuel leak occurred, the color of the health status bar (top bar denoted with a heart) changed from green to yellow with a “FIX LEAK” warning. Participants then had 10 s to click on the bar; otherwise, it would change from yellow to orange and read “FATAL FUEL LEAK” for that specific UAV. A fuel leak occurred 14 times in each scenario.

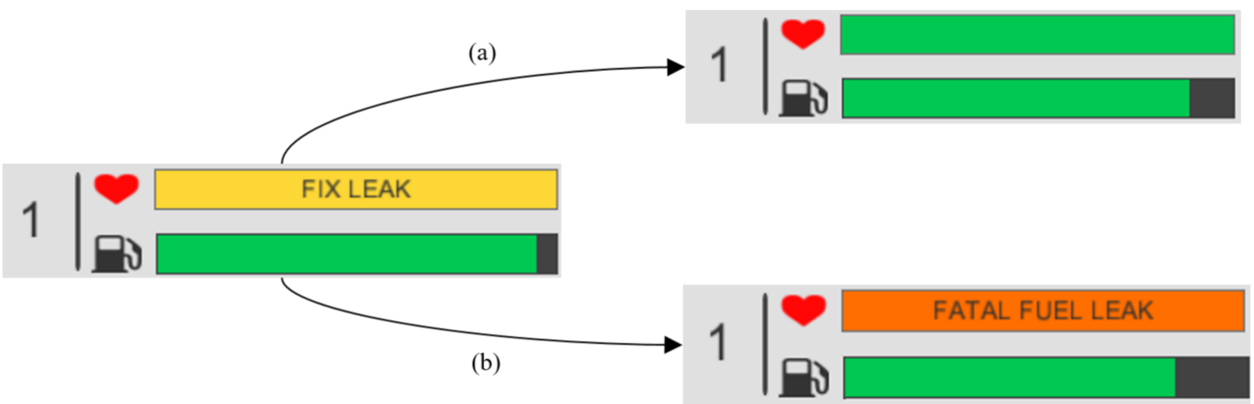


Figure 2.4 Example of how the status of a UAV could change in the General Health panel. Participants were tasked to press the health status bar when a fuel leak occurred: (a) when fuel leak was fixed in time, the health status bar changed from yellow to green and the “FIX LEAK” warning disappeared; (b) when a fuel leak was not fixed in time, the health status bar changed to orange and read “FATAL FUEL LEAK” meaning the fuel leak could no longer be stopped

Chat message task (secondary task)

Participants were tasked with responding to chat messages by selecting between the two options on the left-hand side of the Chat Message panel (Figure 2.5). Responding to chat messages consisted of selecting from one of two options (e.g., selecting ‘yes’ or ‘no’ to yes/no questions). Participants could respond to a chat message until another message appeared and were instructed to accurately answer as many questions as possible. There were 19 chat messages in each scenario.

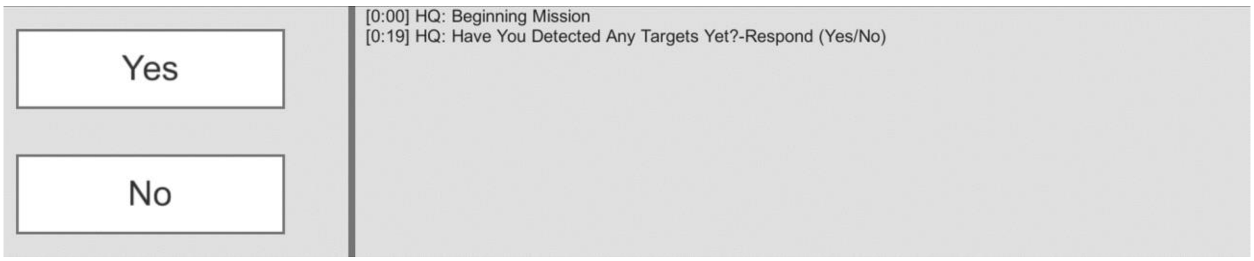


Figure 2.5 Example of an incoming message prompting a response in the Chat Message panel. Participants could select from one of two options for every message

Point system and dependent measures

Table 2.2 shows the point system implemented to encourage participants to prioritize the primary task and avoid task shedding. This point system reinforced the need to prioritize searching for targets in the Video Feed panel, as successfully detecting a target earned the most points. Losing a UAV in the no-fly-zone not only resulted in an immediate loss of points, but also the loss of the opportunity to gain points from that UAV’s target detection task. Whenever the UAV became inoperable, the corresponding UAV video feed became inactive for the remainder for the scenario. The highest scoring participant also earned a bonus $10 gift card. Response time for the primary target detection task was calculated from the appearance of the target to when the participant clicked the “Target” button. For all secondary tasks, response time was calculated from the onset of the event to when the participant responded. Accuracy was calculated as the percentage of correct responses within the time limit for each task.

Table 2.2 Point system for the UAV command and control testbed

Task	Points per Response
Correctly recognizing a target	+100
Correctly recognizing a non-target	+50
All secondary tasks (reroute, fuel leak, and chat message)	+30
Any incorrect or lack of response (false positive or negative to target detection task, UAV flies through no-fly-zone, or “FATAL FUEL LEAK” condition)	-100

Testbed scenarios

Workload was manipulated by varying the number of active UAVs (i.e., highlighted video feeds) in the target detection task. This approach is consistent with previous studies, where workload was manipulated by directly manipulating the load of the primary task (e.g., Hancock et al., 1995). Additionally, this approach was considered appropriate given that the long-term goal of UAV command and control is to increase the load per operator (United States Department of Defense, 2013). The four workload scenarios used in this study are as follows:

1. **Low workload scenario:** There were 3–5 UAVs active for the entirety of the scenario.
2. **High workload scenario:** There were 13–16 UAVs active for the entirety of the scenario.
3. **Medium transitions scenario:** The number of active UAVs increased incrementally. The scenario started at low workload for 20 s, and then one to three active UAVs

were added every 10 s until high workload was reached (13–16 active UAVs). The scenario would remain at high workload for two minutes, before immediately returning to low workload. This cycle repeated five times for this scenario. The solid dark gray line in Figure 2.6 depicts the theoretical number of simultaneously active UAVs over the course of this testbed scenario.

4. **Fast transitions scenario:** The number of active UAVs increased instantaneously. One minute of low workload (3–5 UAVs) was followed by an instantaneous increase to high workload (13–16 UAVs) that lasted for two minutes. After the two minutes of high workload, there was an immediate return to low workload. This cycle repeated five times for this scenario. The dotted light gray line in Figure 2.6 depicts the theoretical number of simultaneously active UAVs over the course of this scenario.

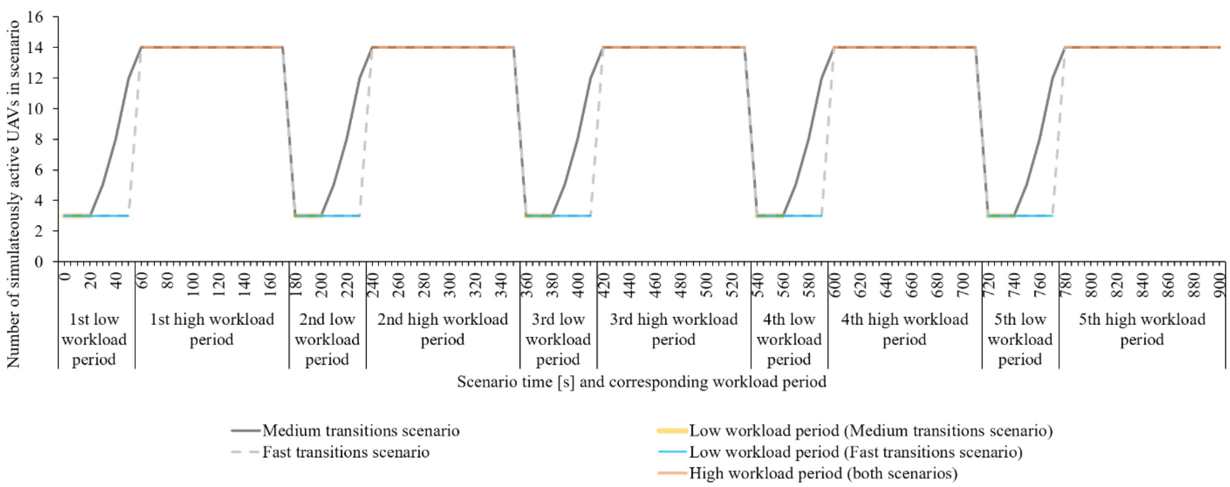


Figure 2.6 The theoretical number of active UAVs throughout the medium and fast transition scenarios. The horizontal axis and denotes workload periods (e.g., 1st low, 1st high, etc.) and is highlighted accordingly

Workload level thresholds (i.e., low and high) were based on pilot testing data using both performance and NASA-TLX measures (Hart & Staveland, 1988). Mean target detection task (primary task) accuracy was approximately 30% higher in low workload compared to high workload. NASA-TLX dimensions of interest included mental demand, temporal demand, and performance. Significant differences between low and high workload were found for all dimensions ($p<.05$; analysis was done with a Friedman test and pairwise comparisons were performed using Mann-Whitney tests). A range of UAVs was used for the low and high workload thresholds because it was possible for the participant to lose a UAV by not rerouting it in time. As such, the range of UAVs allowed the requisite workload level throughout the scenario to be maintained. There was never a situation where the intended workload was not imposed due to the loss of UAVs. Of note, the medium and fast transition scenarios only included transitions from low to high workload, as this transition direction is emblematic of situations likely to occur in data-rich, dynamic domains (e.g., Apollo 12, Murray & Cox, 1989) and was therefore the focus of our work. Also, the high workload periods within the medium and fast transition scenarios were longer than the low workload periods as extended periods of high workload is a challenge of data-rich, dynamic domains (e.g., UAVs command and control; Arrabito et al., 2010; Williams, 2006). In order to have the medium and fast transitions scenarios include the same number of transitions from low to high workload, the time spent in low workload varied slightly between the two scenarios.

Procedures

This research complied with the APA Code of Ethics and was approved by the

Institutional Review Board at Clemson University. Informed consent was obtained from each participant. The experiment took place over two consecutive days at approximately the same time of day. On the first day, participants gave consent and were briefed about the study goals and expectations. Participants then completed a five-minute training session where 4–6 UAVs were always active. By the end of the training session, participants had to demonstrate proficiency by having a minimum accuracy of 70% across all tasks. If not, they could ask questions and reattempt the training session; otherwise, they were excused from the study. Only four participants completed the training session twice and each succeeded on their second attempt. Participants then completed the two constant workload scenarios (low and high workload scenarios) in a randomized order (i.e., 10 participants completed the low workload then high workload scenario and vice versa for the other 11 participants). This approach allowed for the constant workload scenarios to serve as baseline comparisons. On the second day, participants completed the two transitions scenarios (medium and fast transitions scenarios) in a randomized order (i.e., 11 participants completed the medium then fast transitions scenario and vice versa for the other 10). The study lasted about 2.5 hours over two days, with the first day lasting 1.5 hours and the second day lasting 1 hour.

Analysis

Four repeated-measures analysis of variance (RM ANOVA) were performed to address each research question. Of primary interest was to compare primary and secondary task performance, so response time and accuracy of the primary task, i.e., target detection task, was compared to the aggregated response time and accuracy of all the secondary tasks, i.e., reroute, fuel leak, and chat message task.

- Analysis for RQ 2.1 compared performance between the medium and fast transitions scenarios. Specifically, response time and accuracy were analyzed using separate 2×2 RM ANOVA (two testbed scenarios, two task types). Then to analyze the workload history effect on performance, a 2×2×5 RM ANOVA was used. There were two testbed scenarios, (medium, fast), two levels of workload (low, high), and five workload periods (1^{st}, 2^{nd}. . . 5^{th}; refer to the highlighted sections of Figure 2.6 as those were the only sections included in the analysis). Specific main and interaction effects of this RM ANOVA were of interest to our research questions. The dependent performance measures were based on the primary task only, because workload was manipulated using this task and it happened continuously over the course of both scenarios. For this separate analysis, RStudio 1.2.1335 was used with general eta-squared (η_G^2) as the effect size measure for the omnibus test (RStudio Team, 2020). Post-hoc tests with Tukey's adjustment method determined significant differences between means.
- Analysis for RQ 2.2 compared performance in the low workload scenario to low workload of the medium transitions scenario (i.e., the five 20 s low workload periods aggregated) and the fast transitions scenario (i.e., the five 60 s low workload periods aggregated). Specifically, response time and accuracy were analyzed with 3×2 RM ANOVA (three testbed scenarios, two task types).
- Analyses for RQ 2.3 compared performance in the high workload scenario to high workload of the medium and fast transitions scenario (the five 120 s high workload periods aggregated, respectively). Specifically, response time and accuracy were

analyzed with separate 3×2 RM ANOVAs (three testbed scenarios, two task types).

Unless otherwise specified, Bonferroni corrected, Fisher's protected LSD post-hoc tests were performed to test differences between means, and significance was set at α=.05. Before completing the analyses associated with each research question, 2×2×2 (two order types, two transition scenarios, and two task types) mixed ANOVAs were completed to assess whether the testbed scenario order affected primary and secondary task performance. There were no significant main effects of order (p>.05), so analyses for RQ 2.2 and 2.3 proceeded. Violations of normality were assessed prior to analysis, and Greenhouse-Geisser corrections were used when sphericity was violated (using Mauchly's test). SPSS 24 was used for all analyses and partial eta-squared (η_p^2) is reported as a measure of effect size for the omnibus test, and values of .01, .06, and .14 are interpreted as small, medium, and large effect sizes, respectively (Cohen, 1988). For all post-hoc pairwise comparisons, effect sizes were measured using Cohen's d for repeated measures (d_{rm} Lakens, 2013), and values of .2, .5, and .8 are interpreted as small, medium, and large effect sizes, respectively (Cohen, 1988). For all figures, braces indicate groupings of testbed scenarios and brackets with an asterisk denote a significant difference between the testbed scenarios.

Results

RQ 2.1: Medium vs. fast transitions

Response time

When comparing medium to fast transitions, there was no main effect of testbed scenario

(F(1,20)= 1.497, p=.235, η_p^2=.070), but there was a main effect of task type (F(1,20)=110.400, p<.001, η_p^2=.847). The mean response time for the primary task (M=2.85 s, SE=0.03 s) was significantly faster than the secondary tasks (M=3.87 s, SE=0.10 s; p<.001, d_{rm}=2.908). There was no testbed scenario × task type interaction effect (F(1,20)=1.581, p=.223, η_p^2=.073).

To understand the workload history effect of primary task response time during medium and fast transitions, we examine the three-way interaction effect between testbed scenario, workload level, and workload period. Mauchly's test of sphericity indicated that the assumption of sphericity had been violated for the three-way interaction effect (χ^2(9)=0.292, p=.017). Using the Greenhouse-Geisser correction of ε=.608, there was a significant three-way interaction effect (F(2.429, 43.718)=5.123, p=.001, η_G^2=0.049). Post-hoc tests showed that for medium transitions, response times were faster in the 1st low workload period (M=1.98 s, SD=0.34 s) compared to the 3rd – 5th low workload periods (3rd period: M=2.99 s, SD=0.77 s, p<.0001, d_{rm}= 1.627; 4th period: M=2.74 s, SD=1.30 s, p<.001, d_{rm}=.800; 5th period: M=3.06 s, SD=0.70 s, p<.0001, d_{rm}=1.93). In addition, the 2nd low workload period (M=2.23 s, SD=0.76 s) was significantly faster than the 3rd low workload period (p<.001, d_{rm}=1.000) and 5th low workload period (p=.0001, d_{rm}=1.127). For the fast transitions, the 2nd low workload period (M=1.90 s, SD=0.30 s) was significantly faster than the 3rd low workload period (M=2.51 s, SD=0.39 s, p=.022, d_{rm}=1.724). Finally, there was one significant difference between the medium and fast transitions; the 5th low workload period of medium transitions was significantly slower than the 5th low workload period of fast transitions (M=2.28 s, SD=0.43 s, p<.0001, d_{rm}=1.260). Figure 2.7 shows response time over time based on workload period and transition rate.

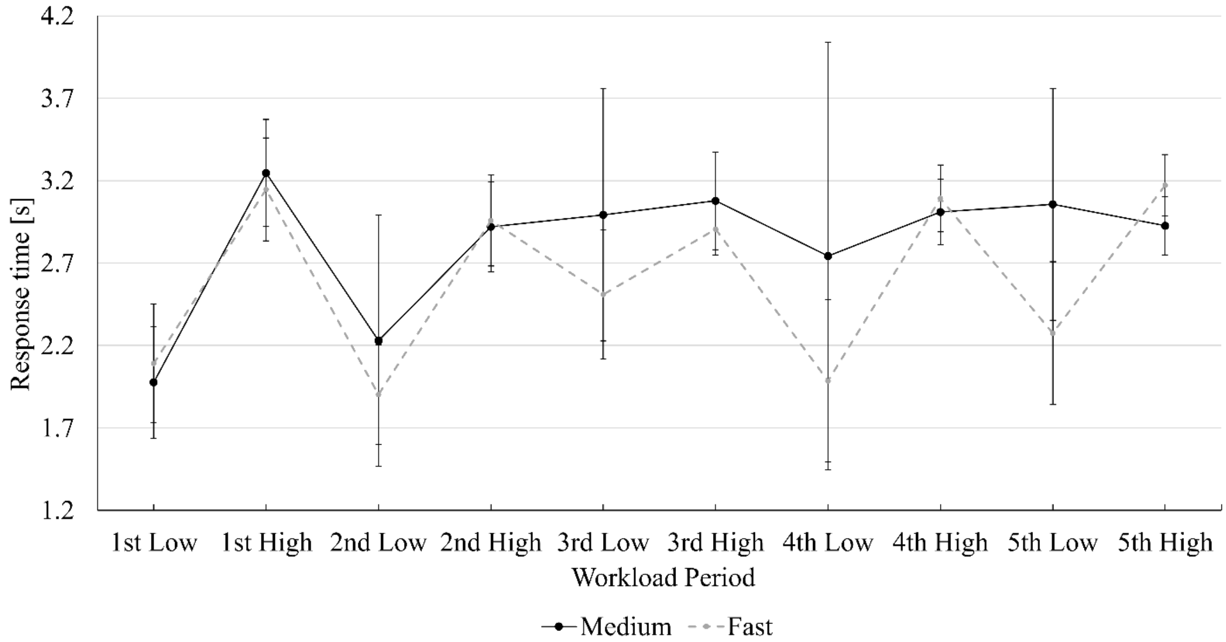


Figure 2.7 Mean primary task response time for each workload period in the medium and fast transitions scenarios. Error bars are standard deviation of the mean

Accuracy

There was no main effect of testbed scenario (F(1,20)=1.552, p=.227, η_p^2=.220), but there was a significant main effect of task type (F(1,20)=579.583, p<.001, η_p^2=.967). There was a significant testbed scenario × task type interaction effect (F(1,20)=6.056, p=.023, η_p^2=.232). For the primary task, the medium transitions scenario (M=69.6%, SE=1.5%) had significantly worse accuracy than fast transitions scenario (M=72.4%, SE=0.8%; p=.048, d_{rm}=.421), but this was not true for the secondary task (Figure 2.8).

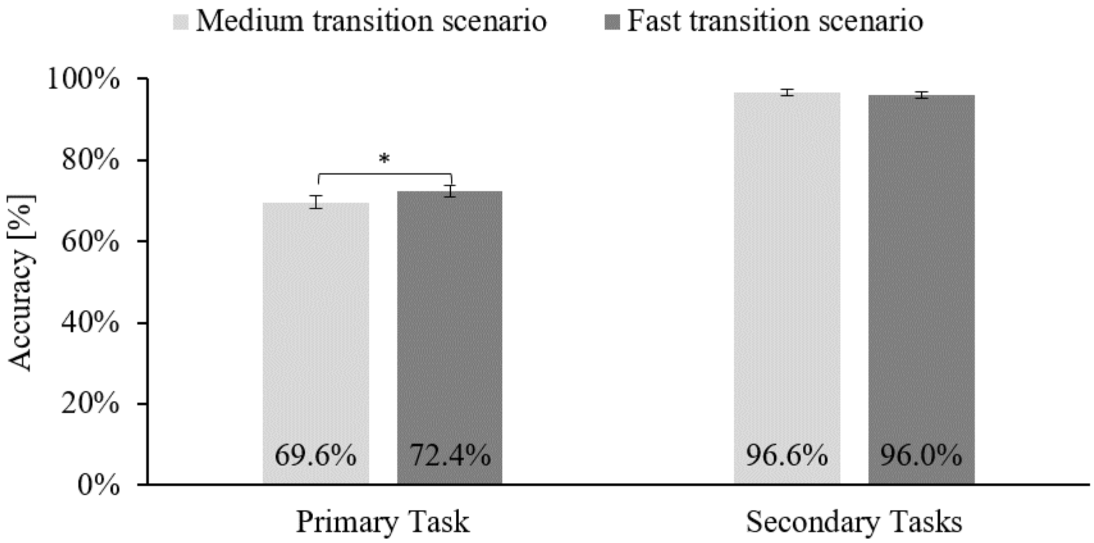


Figure 2.8 Mean accuracy for both task types in the medium and fast transitions scenario. Asterisks (*) denote significant differences between conditions. Error bars represent standard error of the mean

To understand the workload history effect of primary task accuracy during medium and fast transitions, we examine the three-way interaction effect between testbed scenario, workload level, and workload period. There was a significant three-way interaction between testbed scenario, workload level, and period (F(4,72)=3.274, p=.016, η_G^2=.027). Post hoc tests showed that for medium transitions, accuracy for the 1st low workload period (M=99.0%, SD=4.6%) was significantly higher than all subsequent low workload periods (2nd period: M=85.5%, SD=12.7%, p=.008, d_{rm}=1.466; 3rd period: M=85.3%, SD=16.1%, p=.006, d_{rm}=1.208; 4th period: M=87.7%, SD=25.5%, p=.077, d_{rm}=0.638; 5th period: M=80.8%, SD=23.1%, p<.0001, d_{rm}=1.152). With fast transitions, the 4th high workload period (M=74.9%, SD=7.2%) had significantly higher accuracy than 5th high workload period (M=61.6%, SD=9.6%, p=.009, d_{rm}=1.472). Figure 2.9 shows accuracy rates based on workload period and transition rate.

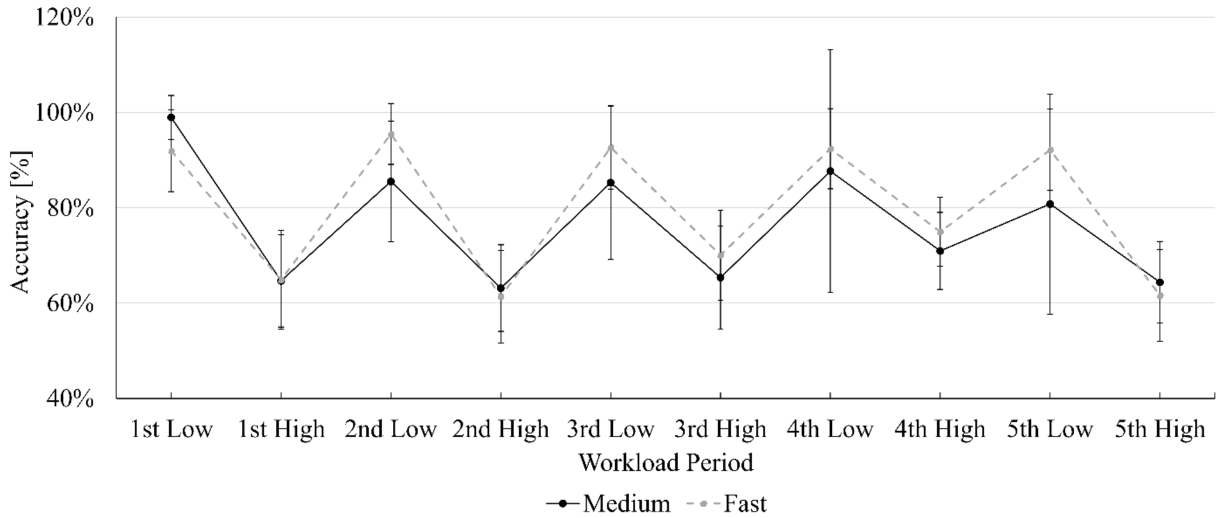


Figure 2.9 Mean primary task accuracy for each workload period in the medium and fast transition scenarios. Error bars are standard deviation of the respective mean

RQ 2.2: Comparing low workload performance

Response time

There was a significant effect of testbed scenario (F(2,40)=20.968, p<.001, η_p^2=.512) and task type (F(1,20)=155.535, p<.001, η_p^2=.886), as well as a significant testbed scenario × task type interaction (F(2,40)=32.238, p<.001, η_p^2=.617). For the primary task, response time during low workload in the medium transitions scenario (*M*=2.64 s, *SE*=0.08 s) was significantly slower than the low workload scenario (*M*=2.38 s, *SE*=0.04 s; *p*<.009, d_{rm}=.844) and low workload in the fast transitions scenario (*M*=2.35 s, *SE*=0.05 s; all *p*<.002, d_{rm}=.917), whereas the latter two did not differ from each other. For the secondary tasks, response time was significantly slower in the low workload scenario (*M*=5.09 s, *SE*=0.26 s) than during low workload in the medium

(M=3.70 s, SE=0.13 s; p<.001, d_{rm}=1.445) and fast transitions scenarios (M=3.58 s, SE=0.17 s; p<.001, d_{rm}=1.509) as seen in Figure 2.10.

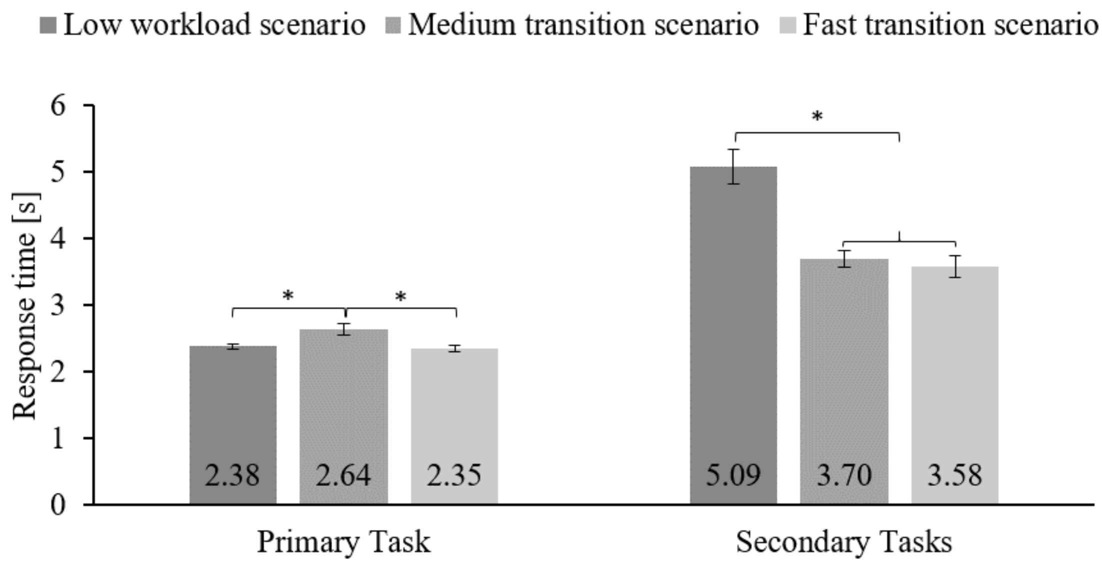


Figure 2.10 Mean response times for both task types during low workload. Asterisks (*) denote significant differences between scenarios. Error bars represent standard error of the mean

Accuracy

Mauchly's test of sphericity indicated that the assumption of sphericity had been violated for workload condition ($\chi^2(2)$=11.732, p=.003). Using the Greenhouse-Geisser correction of ε=.685, there was a significant effect of testbed scenario (F(1.369,27.384)=32.363, p<.001, η_p^2=.618). Across both tasks, mean accuracy for all testbed scenarios were significantly different from each other. The low workload scenario (M=85.8%, SE=1.5%) was significantly lower than low workload in the medium (M=92.7%, SE=1.1%; p=.001, d_{rm}=1.138) and fast transitions scenario (M=95.8%, SE=0.7%; p<.001, d_{rm}=1.591). Accuracy for low workload in the medium transitions scenario was significantly lower than low workload in the fast transitions scenario (p=.022, d_{rm}=.700). There was also a significant effect of task type (F(1,20)=77.774, p<.001, η_p^2=.795), with the mean accuracy for the primary task (M=87.8%, SE=0.9%) being significantly lower than the secondary task (M=95.1%, SE=1.0%; p<.001, d_{rm}=1.671). There was no significant testbed scenario × task type interaction on accuracy F(1.476,29.514)=3.22, p=.068, η_p^2=.139).

RQ 2.3: Comparing high workload performance

Response time

There was a significant effect of testbed scenario (F(2,40)=38.882, p<.001, η_p^2=.660), task type (F(1,20)=109.988, p<.001, η_p^2=.846), as well as a significant testbed scenario × task type interaction (F(2,40)=14.253, p<.001, η_p^2=.416). Primary task response time for the high workload scenario (M=3.38 s, SE=0.06 s) was significantly slower than high workload in the medium (M=3.03 s, SE=0.03 s; p<.001, d_{rm}=1.549) and fast transitions scenarios (M=3.03 s, SE=0.03 s; p<.001, d_{rm}=1.553). For the secondary tasks, response time for the high workload scenario (M=5.22 s, SE=0.22 s) was slower than high workload in the medium (M=4.10 s, SE=0.10 s; p<.001, d_{rm}=1.247) and fast transitions scenarios (M=3.93 s, SE=0.14 s; p<.001, d_{rm}=1.490) as seen in Figure 2.11.

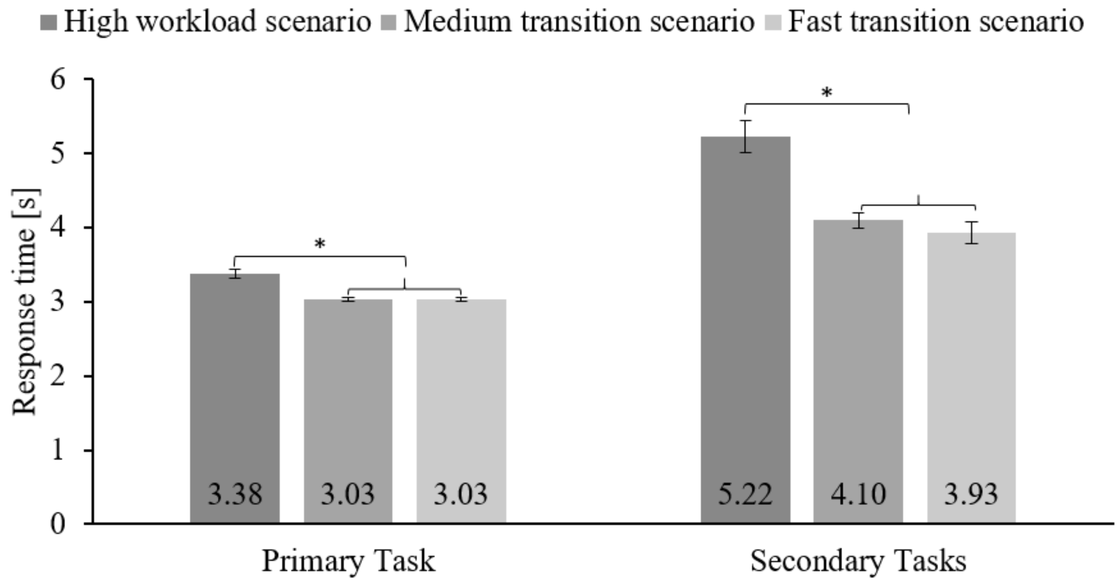


Figure 2.11 Mean response times for both task types during high workload. Asterisks (*) denote significant differences between scenarios. Error bars represent standard error of the mean

Accuracy

There was a significant effect of testbed scenario (F(2,40)=44.028, p<.001, η_p^2=.688). Across both tasks, mean accuracy for the high workload scenario (M=73%, SE=1.3%) was significantly worse than high workload in the medium (M=82.3%, SE=1.0%; p<.001, d_{rm}=1.507) and fast transitions scenarios (M=81.6%, SE=1.0%; p<.001, d_{rm}=1.459). There was also a significant effect of task type (F(1,20)=484.172, p<.001, η_p^2=.960). Mean accuracy for the primary task (M=63.7%, SE=1.3%) was significantly lower than the secondary task (M=94.1%, SE=0.9%; p<.001, d_{rm}=5.777). There was no significant testbed scenario × task type interaction (F(2,40)=3.22, p=.200, η_p^2=.077).

Discussion

Our goal was to better understand how workload transitions influence multitasking performance in a realistic domain, and to assess the applicability of the resource depletion and effort regulation explanations. With respect to RQ 2.1, the analyses found a few, minor differences between medium and fast workload transitions, in that fast transitions had higher primary task accuracy compared to medium transitions overall (Figure 2.8) and faster primary task response time during low workload (Figure 2.10). More notable differences were found when investigating the workload history effect of each scenario (Figure 2.7 and 2.9).

We expected response time and accuracy to degrade and then plateau over time for both transitions scenarios, leading to a transient workload history effect. The results here did reveal a workload history effect, but it was only for low workload periods and it was not the same for medium and fast transitions. Similar to previous work, both transitions scenarios had faster response times in earlier low workload periods, i.e., 1^{st} and 2^{nd}, than later ones, i.e., 3^{rd}–5^{th}. However, the results showed that the workload history effect was more pronounced for medium transitions as response times increased and then plateaued at these slower response times during the latter low workload periods (Bowers et al., 2014; Morgan & Hancock, 2011). For fast transitions, response time increased significantly during the middle of the scenario, i.e., 3^{rd} low workload period, but then recovered to initial speeds later (Jansen et al., 2016; Prytz & Scerbo, 2015). Similarly, for accuracy, we also observed workload history effects that manifested over time during the low workload periods, but only for medium transitions (Moroney et al., 1995). Consistent with previous work and our expectations, accuracy was the highest at the beginning of the medium transitions scenario and then decreased over time (Bowers et al., 2014; Cox-

Fuenzalida & Angie, 2005; Helton et al., 2008; Morgan & Hancock, 2011). On the other hand, with fast transitions, accuracy remained relatively stable over time for low workload periods. The only performance difference within fast transitions was the significant drop in accuracy between the last two high workload periods, although it is unclear if this is due to workload history or fatigue (Cox-Fuenzalida, 2007). Nevertheless, the performance results support a subset of previous work and our expectations: workload history effect are a function of workload transition rate, workload level, and experimental setting, i.e., multitasking during multiple workload transitions, as performance either remained constant, worsened, or fluctuated between worsening and improving over time, depending on these factors. The findings here suggest that dynamic multitasking—the main experimental difference between this line of work and the single-task environment used in Moroney et al. (1995)—may affect workload transition performance differently. Although addressing RQ 2.1 provides further understanding on workload transitions, it does not elucidate the applicability of either the resource depletion or effort regulation explanations. To further understand the applicability of these explanations, comparing low and high performance of medium and fast workload transitions to constant workload is needed.

Regarding RQ 2.2 the majority of findings indicated faster response times and higher accuracy rates during the low workload periods of both the medium and fast transitions scenarios compared to the low workload scenario (e.g., Figure 2.10), which is consistent with some prior work (e.g., Jansen et al., 2016; Krulewitz et al., 1975). The only instance when this was not the case was primary task response time. Specifically, low workload performance during the medium transitions scenario resulted in longer response times compared to low workload performance during the fast transitions scenario and the low workload scenario. Our results are consistent with previous work that found performance improves during low workload periods of workload transitions in the presence of a secondary task (Jansen et al., 2016; Matthews & Desmond, 2002). Our findings also show that fluctuations in workload result in superior performance than when workload is held at a constant low level. In sum, our findings do not support the resource depletion explanation, which predicts that low workload periods of workload transitions will result in worse performance because resources are depleted during high workload periods. It appears that different kinds of multitasking environments may not cause operators to experience resource depletion during a workload transition in the same way (Gluckman et al., 1993). There needs to be more work to examine the impacts of specific contextual factors, as our work suggests that diversified task demands can help thwart resource depletion effects in complex work environments.

For RQ 2.3, we found that primary and secondary task performance was better during the high workload periods of the medium and fast transitions scenarios, compared to the high workload scenario (e.g., Figure 2.11). However, this improvement in performance is in contrast with some previous work that has found performance decrements during high workload of workload transitions (Cox-Fuenzalida, 2007; Cox-Fuenzalida & Angie, 2005; Gluckman et al., 1993; Krulewitz et al., 1975; Moroney et al., 1995). Previous work did not include secondary tasks and/or multiple workload transitions. The results here suggest that the effort regulation explanation, which suggests that participants can effectively redistribute their mental resources as task demands change, may be dependent not only on the presence of a workload transition, but also multiple workload transitions—the main difference in this study that sets it apart from previous ones. Further research, especially studies that modulate the number of workload transitions, is needed to corroborate this finding, and would shed further understanding on how to improve performance during high workload (Grier et al., 2008).

Overall, the findings here demonstrate that existing theoretical explanations cannot fully explain the effects of workload transitions on performance. Instead, our work further supports the nuanced nature of workload transitions, as we explored contexts that have been overlooked to-date. For instance, existing explanations do not distinguish between primary and secondary task performance. Our results showed this to be an important consideration, since secondary task performance improved with both types of workload transitions, but this was not always the case for the primary task. This finding highlights the importance of considering the role of secondary tasks in complex domains, because it has been historically assumed that primary task performance would be prioritized over secondary task performance (Wickens et al., 2015). In reality, task prioritization may vary depending on context, especially in dynamic environments such as UAV operations where operators are juggling multiple, interrelated tasks and responsibilities (Jansen et al., 2016; Matthews et al., 1996).

Our work demonstrates this notion empirically, as participants attended successfully to secondary tasks, even if at times it led to an immediate cost to primary task performance. This satisficing approach supports previous work that has shown that participants may change how they prioritize tasks as they recognize task interdependencies can change over time (Jansen et al., 2016). Therefore, it is important for future work to investigate workload transitions with interdependent, dynamic tasks to fully understand: (a) the effect of workload transitions on performance; (b) the applicability of existing theoretical explanations; and (c) contextual factors from different occupational environments. Overall, our results suggest that occupational factors in these complex and dynamic domains impact workload transition performance in different ways than expected and the findings can be used to inform design. For example, it may be beneficial to strategically engage and reengage operators in UAV command and control with diverse tasks, so they can manage mental resources more effectively over time, just as the workload history analysis suggested. Such a strategy may help negate the effects due to vigilance decrements during constant low workload and data overload during constant high workload. As a result, performance may improve because mental resources could be more effectively employed. This strategy could be taken into consideration occupational and technology design.

Limitations

Our results show workload transition performance is nuanced. One limitation of the current work is that only two workload transition rates were examined. Transition rate was not thoroughly explored, because it would have affected either the number of transitions per scenario if scenario length was held constant, or the lengths of each scenario if the number of transitions was held constant. Although this presents a limitation of this work, the subsequent analyses and discussion are unaffected because our overarching goal was to examine how the rate of workload transitions affects performance during the low and high workload periods. Nevertheless, it may be of interest to consider different rates in future work as the performance differences currently observed between transition rates did begin to inform theory applicability. While outside of the scope of this study, scenario duration is another limitation worth mentioning as there is evidence that workload transitions may have longer-term ramifications (Cox-Fuenzalida, 2007). Another potential limitation was that low and high workload thresholds were set equivalent across participants. As a result, high workload may have been deemed to be more difficult for some participants than others (Prytz & Scerbo, 2015). Future work could individually tailor workload thresholds to account for individual differences, however such an experimental design comes with its own challenges (see Bowers et al., 2014; McKendrick & Harwood, 2019). It may be more advantageous for future work to compare workload history effects between subgroups of performers (e.g., best and worst performers). Given the spread of the data for both the performance measures was considerable, it warrants further explanation at an individual level. Other occupationally-relevant effects, such as task type (e.g., manual vs. verbal tasks), expectancy effects (e.g., Landman et al., 2017), and individual differences, (e.g., expertise and personality; Cox-Fuenzalida, Angie, et al., 2006; Cox-Fuenzalida et al., 2004) should be explored to evaluate the present findings and provide further details for these domains. Similarly, it may be beneficial to include real-time, unobtrusive measures of the operator during workload transitions as a means to capture his/her real-time state. For example, measuring how the operator allocates his/her visual attention to features of the environment is informative on how the operator is managing the tasks beyond just speed and accuracy. This may better explain the nuanced performance trends and potentially the applicability of current theoretical explanations surrounding workload transitions (Duchowski, 2017). This will be particularly critical as the frequency of overt actions from the operator decreases given the increased reliance on automation and autonomy in these environments (Cummings et al., 2019; Sibley et al., 2015).

Conclusion

This study highlights the nuanced nature surrounding workload transitions, specifically that the effects evolve differently over time in dynamic, multitasking domains. This study addressed the need to consider different transition rates, multitasking, and multiple transitions to understand the effects of workload transitions overall and over time (Jansen et al., 2016). Our results showed some differences in performance between medium and fast transitions, but transitions in general resulted in faster and more accurate performance than constant workload. In total, our findings provide further support for the effort regulation explanation (Hockey, 1997); however, future work should investigate the applicability of existing theoretical explanations by better understand how the operator is managing the workload transition (Edwards et al., 2017) and as they relate to different occupational factors and settings. Although our findings have implications on the design of systems for operators in various complex domains, future work needs to address how to best integrate them to ensure operators can safely cope with workload transitions, such as the operator's management approach and other aspects of multi-UAV environments (Cummings et al., 2019; Ramchurn et al., 2015; Sibley et al., 2015).

References

Arrabito, G. R., Ho, G., Lambert, A., Rutley, M., Keillor, J., Chiu, A., AU, H., & Hou, M. (2010). *Human factors issues for controlling uninhabited aerial vehicles*. Defence Research and Development Canada Toronto. www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA543186

Baker, A. L., & Keebler, J. R. (2017). Factors affecting performance of human-automation teams. In P. Savage-Knepshield & J. Chen (Eds.), *Advances in Human Factors in Robots and Unmanned Systems* (pp. 331–340). https://doi.org/10.1007/978-3-319-41959-6

Bowers, M. A., Christensen, J. C., & Eggemeier, F. T. (2014). The effects of workload transitions in a multitasking environment. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 220–224. https://doi.org/10.1177/1541931214581046

Clare, A. S., Hart, C. S., & Cummings, M. L. (2010). Assessing operator workload and performance in expeditionary multiple unmanned vehicle control. *48th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition*, 1–10. https://doi.org/10.2514/6.2010-763

Cohen, J. (1988). Analysis of variance. In Statistical power analysis for the behavioral sciences (pp. 273–406).

Cox-Fuenzalida, L.-E. (2007). Effect of workload history on task performance. Human Factors, *49*(2), 277–291. https://doi.org/10.1518/001872007X312496

Cox-Fuenzalida, L.-E., & Angie, A. D. (2005). The effects of workload history on dual task performance. Current Psychology, *24*(3), 171–179. https://doi.org/10.1007/s12144-005-1020-y

Cox-Fuenzalida, L.-E., Angie, A., Holloway, S., & Sohl, L. (2006). Extraversion and task performance: A fresh look through the workload history lens. Journal of Research in Personality, *40*(4), 432–439. https://doi.org/10.1016/j.jrp.2005.02.003

Cox-Fuenzalida, L.-E., Beeler, C., & Sohl, L. (2006). Workload history effects: A comparison of sudden increases and decreases on performance. Current Psychology, *25*(1), 8–14. https://doi.org/10.1007/s12144-006-1012-6

Cox-Fuenzalida, L.-E., Swickert, R., & Hittner, J. B. (2004). Effects of neuroticism and workload history on performance. Personality and Individual Differences, *36*(2), 447–456. https://doi.org/10.1016/S0191-8869(03)00108-9

Cummings, M. (2014). Man versus machine or man + machine? *IEEE Intelligent Systems*, *29*(5), 62–69.

Cumming, R. W., & Croft, P. G. (1973). Human information processing under varying task demand. *Ergonomics*, *16*(5), 581–586. https://doi.org/10.1080/00140137308924548

Cummings, M. L., Huang, L., Zhu, H., Finkelstein, D., & Wei, R. (2019). The impact of increasing autonomy on training requirements in a UAV supervisory control task. *Journal of Cognitive Engineering and Decision Making*, *13*(4), 295–309. https://doi.org/10.1177/1555343419868917

Department of Defense. (2017). *Unmanned systems integrated roadmap FY 2017-2042*.

Devlin, S. P., Byham, J. K., & Riggs, S. L. (2021). Does what we see shape history? Examining workload history as a function of performance and ambient/focal visual attention. *Transactions on Applied Perception*, *8*(2), 1-17. https://doi.org/10.1145/3449066

Devlin, S. P., Moacdieh, N. M., Wickens, C. D., & Riggs, S. L. (2020). Transitions between low and high levels of mental workload can improve multitasking performance. *IISE*

Transactions on Occupational Ergonomics and Human Factors, *8*(2), 72–87. https://doi.org/10.1080/24725838.2020.1770898

Duchowski, A. T. (2017). *Eye tracking methodology: Theory and practice* (3rd ed). Springer International Publishing AG.

Edwards, T., Gabets, C., Mercer, J., & Biener, N. (2017). Task demand variation in air traffic control: Implications for workload, fatigue, and performance. In *Advances in Human Aspects of Transportation* (Issue January, pp. 91–102). Springer. https://doi.org/10.1007/978-3-319-41682-3

Feitshans, G., Rowe, A., Davis, J., Holland, M., & Berger, L. (2008). Vigilant spirit control station (VSCS) the face of counter. AIAA Guidance, Navigation and Control Conference and Exhibit. https://doi.org/10.2514/6.2008-6309

Gluckman, J. P., Warm, J. S., Dember, W. N., & Rosa, R. R. (1993). Demand transitions and sustained attention. The Journal of General Psychology, *120*(3), 323–337. https://doi.org/10.1080/00221309.1993.9711151

Grier, R. A., Wickens, C. D., Kaber, D., Strayer, D., Boehm-Davis, D., Trafton, J. G., St. John, M. (2008). The red-line of workload: Theory, research, and design. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1204–1208. https://doi.org/10.1177/154193120805201811

Hancock, P. A., Williams, G., Manning, C. M., & Miyake, S. (1995). Influence of task demand characteristics on workload and performance. *The International Journal of Aviation Psychology*, *5*(1), 63–86. https://doi.org/10.1207/s15327108ijap0501_5

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, 52, 139–183.

https://doi.org/10.1016/S0166-4115(08)62386-9

Helton, W. S., Shaw, T., Warm, J. S., Matthews, G., & Hancock, P. (2008). Effects of warned and unwarned demand transitions on vigilance performance and stress. *Anxiety, Stress, and Coping*, *21*(April), 173–184. https://doi.org/10.1080/10615800801911305

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive energetical framework. *Biological Psychology*, *45*, 73–93. https://doi.org/10.1016/S0301-0511(96)05223-4

Hooey, B. L., Kaber, D. B., Adams, J. A., Fong, T. W., & Gore, B. F. (2017). The underpinnings of workload in unmanned vehicle systems. *IEEE Transactions of Human-Machine Systems*, *99*, 1–16. https://doi.org/10.1109/TCST.2007.894653

Huey, B. M., & Wickens, C. D. (Eds.). (1993). *Workload transition: Implications for individual and team performance.* National Research Council.

Jansen, R. J., Sawyer, B. D., van Egmond, R., de Ridder, H., & Hancock, P. A. (2016). Hysteresis in mental workload and task performance: The influence of demand transitions and task prioritization. *Human Factors*, *58*(8), 1143–1157. https://doi.org/10.1177/0018720816669271

Kim, N. Y., House, R., Yun, M. H., Nam, C. S., & Harrivel, A. R. (2019). Neural correlates of workload transition in Multitasking: An ACT-R model of hysteresis effect. *Frontiers in Human Neuroscience*, *12*(January), 1–14. https://doi.org/10.3389/fnhum.2018.00535

Krulewitz, J. E., Warm, J. S., & Wohl, T. H. (1975). Effects of shifts in the rate of repetitive stimulation on sustained attention. *Perception & Psychophysics*, *18*(4), 245–249. https://doi.org/10.3758/BF03199369

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(863), 1–12. https://doi.org/10.3389/fpsyg.2013.00863

Landman, A., Groen, E. L., van Paassen, M. M., René., Bronkhorst, A. W., & Mulder, M. (2017). Dealing with unexpected events on the flight deck: A conceptual model of startle and surprise. *Human Factors*, *59*(8), 1161–1172. https://doi.org/10.1177/0018720817723428

Matthews, M. L. (1986). The influence of visual workload history on visual performance. *Human Factors*, *28*(6), 623–632. https://doi.org/10.1177%2F001872088602800601

Matthews, G., & Desmond, P. A. (2002). Task-induced fatigue states and simulated driving performance. *The Quarterly Journal of Experimental Psychology*, *55A*(2), 659–686. https://doi.org/10.1080/02724980143000505

Matthews, G., Sparkes, T. J., & Bygrave, H. M. (1996). Attentional overload, stress, and simulated driving performance. *Human Performance*, *9*(1), 77–101. https://doi.org/10.1207/s15327043hup0901_5

McKendrick, R., & Harwood, A. (2019). Cognitive workload and workload transitions elicit curvilinear hemodynamics during spatial working memory. *Frontiers in Human Neuroscience*, *13*(405), 1–16. https://doi.org/10.3389/fnhum.2019.00405

Morgan, J. F., & Hancock, P. A. (2011). The effect of prior task loading on mental workload: An example of hysteresis in driving. *Human Factors*, *53*(1), 75–86. https://doi.org/10.1177/0018720810393505

Moroney, B. W., Warm, J. S., & Dember, W. N. (1995). Effects of demand transitions on vigilance performance and perceived workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1375–1379. https://doi.org/10.1177%2F154193129503902104

Murray, C. A., & Cox, C. B. (1989). Apollo: The race to the moon. New York: Simon & Schuster.

Prytz, E. G., & Scerbo, M. W. (2015). Changes in stress and subjective workload over time following a workload transition. *Theoretical Issues in Ergonomics Science, 16*(6), 586–605. https://doi.org/10.1080/1463922X.2015.1084397

Ramchurn, S. D., Fischer, J. E., Ikuno, Y., Wu, F., Flann, J., & Waldock, A. (2015). A study of human-agent collaboration for multi-UAV task allocation in dynamic environments. *IJCAI International Joint Conference on Artificial Intelligence*, 1184–1192. Buenos Aires, Argentina.

RStudio Team (2020). *RStudio: Integrated Development for R. RStudio*, PBC, Boston, MA http://www.rstudio.com/

Sibley, C., Coyne, J., & Morrison, J. (2015). Research considerations for managing future unmanned systems. *AAAI Spring Symposium*, 54–59.

Ungar, N. R., Matthews, G., Warm, J. S., Dember, W. N., Thomas, J. K., Finomore, V. S., & Shaw, T. H. (2005). Demand transitions and tracking performance efficiency: Structural and strategic models. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1523–1526. https://doi.org/10.1177/154193120504901704

Wickens, C. D., Gutzwiller, R. S., & Santamaria, A. (2015). Discrete task switching in overload: A meta-analyses and a model. *International Journal of Human Computer Studies, 79*, 79–84. https://doi.org/10.1016/j.ijhcs.2015.01.002

Wickens, C. D., Tsang, P., & Pierce, B. (1985). The dynamics of resource allocation. *Advances in Man-Machine Systems Research, 2*, 1–49.

Williams, K. W. (2006). Human factors implications of unmanned aircraft accidents: Flight-control problems. https://doi.org/10.1016/S1479-3601(05)07008-6

CHAPTER 3

What do scan-based eye tracking metrics inform about workload transition performance?

Introduction

To better understand the performance trends of Chapter 2 and how they contribute to both the theory and applicability of workload transitions, Chapter 3 examines the user's visual attention allocation patterns. Previous workload transition research has also included non-performance measures of the operator like, his/her personality (Cox-Fuenzalida, Angie, et al., 2006; Cox-Fuenzalida et al., 2004) or subjective experience (Fallahi et al., 2016; Helton et al., 2008; Jansen et al., 2016; Morgan & Hancock, 2011). These measures have shed more light on various aspects of workload transitions, but their limitations may hinder their overall applicability in operational settings, e.g., the need to interrupt the operator while they complete tasks in the environment, (Estes et al., 2015; Matthews et al., 2019; McKendrick & Cherry, 2018). Physiological responses avoid some these shortcomings and have found to have some preliminary use for workload transitions (Bowers et al., 2014; Kim et al., 2019). Several studies have used eye tracking techniques to understand differences between low and high workload (see Coral, 2016 for a review), but there is little work exploring how the operator visually attends to tasks as workload transitions.

The goal of the present work is to determine whether eye tracking analysis can be a non-invasive and quantitative means to further understand how operators are impacted by workload transitions. Nine scan-based metrics were calculated and compared across the four testbed scenarios in Chapter 2 and one metric was calculated across the low and high workload periods of the medium and fast transitions scenarios because it considers the sequence of fixations and saccades as they happen over time in its calculation (Krejtz et al., 2016). Understanding visual attention allocation patterns can be an indication of how the operator is managing workload transitions—for instance, does the operator consistently concentrate their visual attention or disperse it across the entire display? Knowing this information can potentially explain the performance trends observed in Chapter 2 and ideally inform the design of environments where workload transitions.

Previous use of psychophysiological measures in workload and workload transitions

To date, workload transition research has rarely included eye tracking, but some research has included other psychophysiological measures (Bowers et al., 2014; Kim et al., 2019). For example, electroencephalograms (EEG) have shown an increase in the electrical activity in several cognitive-related areas of the brain during workload transitions. However, there are unexplained inconsistencies with some EEG correlates, especially when considering the performance trends over time or the correlates of constant workload conditions (Bowers et al., 2014; Kim et al., 2019). It is worth investigating if eye tracking can serve as a less invasive and less complex psychophysiological measure to capture the impacts of workload transitions in real-time. Of specific interest is to understand how operators use mental resources to manage complex environments and understand how it impacts performance (Edwards et al., 2017).

Unlike workload transitions, different types of psychophysiological measures have been extensively explored when studying workload, including eye tracking (see reviews in Cain, 2007; Young et al., 2015). The most frequently used eye tracking measures for workload evaluation are pupillometry metrics such as pupil diameter (Hampson et al., 2010) and eye blink frequency and duration (Hwang et al., 2008; Veltman & Gaillard, 1996). These measures are often positively correlated with workload; however, such measures are also sensitive to other extraneous factors, such as the amount of light in the environment (Coral, 2016; Monfort et al., 2016). Although avoiding these problems is possible (e.g., Duchowski et al., 2020; Rozado & Dunser, 2015), there are simpler, and potentially more advantageous alternatives, such as scan-based metrics, i.e., ones that capture how an individual is viewing the display (Poole & Ball, 2006).

Scan-based metrics inform how an individual is extracting and sampling the visual information to manage the current environment, which is a present need for both workload transition theory and design. For example, some of the previous research finds high workload leads to the range of attention to narrow, the size of eye movements to increase, and attention on specific items to last longer (Rantanen & Goldberg, 1999; Savage et al., 2013). Questions that scan-based metrics can answer about workload transitions include: *In comparison to constant workload, how dispersed was visual attention overall? How large were eye movements across the display? How long did attention last? How were the multiple tasks attended to?* By answering such questions, scan-based metrics can provide insight into the participant's visual attention allocation patterns during workload transitions, which helps explain the performance trends of Chapter 2.

Fixations and saccades are the basis of most scan-based eye tracking metrics, i.e., the way in which visual attention patterns are quantified (Poole & Ball, 2006). The selected scan-based metrics, with their respective definitions, are summarized in Table 3.1. Some of these metrics have been extensively used in workload research (e.g., fixation duration and saccade amplitude; De Rivecourt et al., 2008; Di Stasi et al., 2013), whereas other measures have been valuable for other research topics in realistic environments. For example, stationary gaze entropy and gaze transition entropy measure the randomness of an individual's attention transitions across the set of AOIs and has informed task completion strategies in simulated aviation, surgical, and driving environments (Di Stasi et al., 2016; Shiferaw et al., 2018; Shiferaw et al., 2019).

Moacdieh and Sarter (2015) classified scan-based metrics in to three distinct categories: spread (*where are users looking?*), directness (*how efficiently are users scanning?*), and duration (*how long are users looking at a certain area?*). Spread metrics have rarely been explored in the context of workload (e.g., Rantanen & Goldberg, 1999) and only a small selection of directness metrics, such as mean saccade amplitude (e.g., Savage et al., 2013) have been previously explored in this context. Both spread and directness measures have been informative on *how* visual attention is allocated during the task, such as: *Was attention evenly distributed in the environment? Did attention reach a certain area efficiently or was there a lot of inefficient back-and-forth scanning?* Duration measures have been successful in discriminating low and high workload, but their ability to do so among multiple, interrelated tasks has rarely been tested. Given the testbed places tasks across distinct areas on the display, answering the aforementioned questions will inform the effectiveness of the task completion strategies used in Chapter 2.

Table 3.1 Scan-based eye tracking metrics investigated in Chapter 3

Metric	**Definition and Calculation**
Spread metrics (where are users generally looking?)	
Convex hull area [pixels^2]	The minimum convex area which contains the fixation points (Goldberg & Kotval, 1999). This is calculated using the Matlab function convHull, with the X and Y positions of the fixation points as input. The maximum area of the screen is 2,560×1600 = 4.096 x 10^6 pixels^2 A larger convex hull area indicates more spread of gaze points and larger cognitive load as the user attempts to sample all the information available within the display (Di Nocera et al., 2007) However, note that there are opposing views related to task load and the visual field of view (Coral, 2016)
Spatial density	The number of grid cells containing gaze points divided by the total number of cells (Goldberg & Kotval, 1999). A 20×20 evenly-divided grid (128×80 pixels per cell) was created to cover the full screen dimensions. Similar to convex hull area, a higher spatial density would indicate a larger dispersion of attention.
Stationary gaze entropy (SGE)	Stationary gaze entropy indicates how equally distributed a person's attention is, with larger values indicating more evenly spread attention across areas of interest and lower values indicating more narrowed attention (Krejtz et al., 2015). It is calculated using the following equation $$H_s = -\sum_{i \in AOIs} p_i \log_2 p_i$$ where p_i represents the proportion of transitions to the i^{th} state, i.e., the i^{th} AOI (the AOIs are as defined in Figure 2.1) from on all the state transitions based on the Markov property (i.e., transitions to a given state only depend on the current state; Shiferaw et al., 2019).
Directness metrics (how purposeful are attention transitions?)	
Mean saccade amplitude [pixels]	The average distance traveled during a saccade (Smeets & Hooge, 2003). Higher mean saccade amplitude indicates lower scanning efficiency (Gegenfurtner et al., 2011).
Scanpath length per second [pixels/s]	The sum of all the saccade lengths divided by the total time. Similar to mean saccade amplitude, a larger scanpath length indicates less efficiency (Goldberg & Kotval, 1999).
Backtrack rate [/s]	A backtrack is defined as an angle between two saccades that is greater than 90º (Goldberg & Kotval, 1999), indicating a change in direction. A higher backtrack rate indicates lower efficiency.
Gaze transition rate [grid cells/s]	The rate of transitions between equal grid cells (Goldberg & Kotval, 1999). A higher rate of transitions indicates lower efficiency. The same grid cells used for spatial density were used here.
Gaze transition entropy (GTE)	Gaze transition entropy represents the randomness and complexity of a person's eye movements, with higher values indicating more randomness and lower efficiency (Krejtz et al., 2015). It is calculated based on the following formula: $$H_t = -\sum_{i \in AOIs} p_i \sum_{j \in AOIs} p_{ij} \log_2 p_{ij}$$ where p_i is as described in stationary gaze entropy, and p_{ij} is the probability of transitioning form state *i* to state *j* in one fixation. Assuming the Markov property holds, this was calculated by counting the number of transitions from *i* to *j* and then dividing by the total number of transitions from *i*

	(Shiferaw et al., 2019). This was done for each pairing of AOIs (the AOIs are as defined in Figure 2.1).
Duration metric (how long, in general, does attention last?)	
Fixation duration [ms]	The amount of time a fixation lasts. A lower mean fixation duration suggests the user is extracting information quickly (Jacob & Karn, 2003).

Using eye tracking techniques to understand workload history

Often, metrics are aggregated across experimental conditions for comparison, which overlooks how visual attention patterns change over time (Cutrell & Guan, 2007; Goldberg & Kotval, 1999; Jarodzka et al., 2010). For example, a study conducted by Jiang et al. (2014) found that as participants completed a web search task, their performance and scan patterns were inversely proportional to the workload changes over time. In other words, as participants completed more search tasks, both the likelihood of selecting the correct search result and mean fixation duration decreased, demonstrating the importance of capturing the evolution of visual attention as it happens. However, the interpretation of a certain metric's trend over time can be convoluted (Pan et al., 2004). Advanced metrics able to account for the order in which certain types of fixations and saccades occur over time have the potential to capture and accurately depict the evolution of visual attention. The metric, coefficient K, is a viable candidate to better understand the workload history effects observed in Chapter 2 (Krejtz et al., 2016).

Coefficient *K*: A dynamic measure of ambient and focal visual attention

Coefficient K accounts for changes in the magnitude and sequence of each fixation and saccade in a scan pattern. It has effectively distinguished between the two types of visual attention used during visual search tasks: ambient and focal (Krejtz et al., 2016). Ambient visual attention occurs when people are gaining spatial orientation (i.e., “getting a sense” of the environment), whereas focal visual attention occurs when people are processing the details of the environment (Buswell, 1935; Previc, 1998). Ambient visual attention usually consists of a pattern where sequences of short fixations are followed by long saccades. Oppositely, focal visual attention usually consists of sequences of long fixations followed by short saccades (Buswell, 1935; Velichkovsky et al., 2005). The interaction between ambient and focal visual attention in scene perception is dynamic (Velichkovsky et al., 2005; Previc, 1998). For example, when a scene is initially being examined, there is typically more ambient visual attention (i.e., shorter fixations and longer saccades), but as objects are identified, there is more focal visual attention (i.e., fixation durations increase and saccades decrease; (Irwin & Zelinsky, 2002; Over et al., 2007).

When it comes to studying visual attention, most analyses aggregate the duration and magnitude of the fixations and saccades over time. There has been limited work analyzing how scan-based metrics change over time and their relation to overall changes in visual attention patterns. Pannasch et al. (2008) compared fixation durations and saccade amplitudes during the early and late phases of scene perception to better understand the relationship between ambient and focal visual attention over time. However, their analysis aggregated fixations and saccades over blocks of time so it did not capture how the sequence of fixations and saccades impacts the evolution of visual attention patterns over time. Krejtz et al. (2016) developed a metric that distinguishes between ambient and focal visual attention each time a fixation or saccade occurs. Equation (3.1) shows how coefficient K is the difference between standardized values (Z-score) of each fixation duration d_i and its following saccade amplitude (a_i+1):

$$K_i = \frac{d_i - \mu_d}{\sigma_d} - \frac{a_i - \mu_a}{\sigma_a} \text{ such that } \frac{1}{n}\sum_{i=1}^{n} K_i = 1 \tag{3.1}$$

Here, μ_d is the mean fixation duration, μ_a is the mean saccade amplitude, σ_d and σ_a are the respective standard deviations. These parameters are for the entire data set to account for any bias. Coefficient *K* is a measure of standard deviation; a value of 1 indicates that "the duration of the current fixation is beyond 1 standard deviation longer than the subsequent saccade amplitude," whereas a value of -1 indicates that "a saccade is more than 1 standard deviation longer than the preceding fixation duration" (Krejtz et al., 2016). Positive coefficient *K* values are an indicator of focal attention as they occur when long fixations are followed by short saccades. Negative coefficient *K* values are an indicator of ambient visual attention as they occur when short fixations are followed by long saccades (Velichkovsky et al., 2005). A coefficient *K* value that approaches zero suggests that fixations and subsequent saccades are relatively equal with their respective means and is not an indicator of either attention type. (*Note: This occurrence is rare on an individual level, although may occur when averaging). Previous research has used it to distinguish the visual attention patterns when viewing artwork (Krejtz et al., 2016), completing cartographic tasks (Krejtz et al., 2017), and between socially anxious and non-anxious viewers (Krejtz et al., 2018). In all cases, coefficient *K* has been more informative on the details of the focal-ambient viewing dynamics, which further explains the viewer's overall visual attention patterns in these environments. Coefficient *K* may further shed light on the effects of workload history by quantifying the visual attention patterns in dynamic environments.

Motivation

This work aims to further understand the performance effects of workload transitions by using scan-based eye tracking metrics to assess how operators relied on their visual attention with different workload transition rates and over time. The goal is to gain perspective on how the participants relied on mental resources during workload transitions, while also potentially informing the design of intelligent technology for domains where workload transitions are prevalent. We are exploring the potential of the aforementioned metrics because they characterize real-time visual attention in a straightforward format and allow for direct interpretation. This chapter examines whether (a) visual attention type is informative of any of the present workload history effects and (b) scan-based metrics can discern between transitioning and constant workload and shed light on their respective performance trends. Specifically, the research questions are:

1. **RQ 3.1:** Is coefficient *K* informative of the workload history effects observed during medium and fast transitions?
    a. **Expectations:** Based on its previous success (Krejtz et al., 2017; Krejtz et al., 2018), we expect coefficient *K* to serve as a quantitative measure of people assessing the current needs of the environment and then developing effective management strategies based on that assessment, i.e., the effort regulation explanation (Hockey, 1997). If this is the case, then participants will first survey the overall environment, i.e., use ambient visual attention which produces negative coefficient *K* values, to “get a sense” and evaluate its dynamics. During this evaluation, performance may suffer if ambient visual attention is not ideal for the current environment. However, once participants develop a management strategy to account for varying workloads, coefficient *K* values will most likely be positive during high workload periods, i.e., focal

visual attention, and negative during low workload periods, i.e., ambient visual attention.

2. **RQ 3.2:** Are scan-based metrics informative on the performance differences between transitioning and constant workload?
    a. **Expectations:** Performance will worsen when participants distribute their visual attention to many, wide-ranging areas of the display and when they do not transition efficiently between those different areas (Goldberg & Kotval, 1999; Shiferaw et al., 2019). We expect that performance decrements will coincide with increased values of spread and duration metrics and decreased values of directness metrics.

Combining the findings from Chapter 2 and 3 leads to a comprehensive, initial investigation of the effects of transition rate on multitasking performance and visual attention allocation patterns.

Method

Participants

The eye tracking data of the participants from Chapter 2 was analyzed for the present research goals. All participants self-reported normal or corrected-to-normal vision.

Experimental setup

The same experimental setup from Chapter 2 was used for the present research goals. A desktop-mounted corneal reflection FOVIO eye tracker (Seeing Machines platform) was used to collect gaze data (sampling rate (fs) of 60 Hz, reported mean degree of error is 0.78 with a standard deviation of 0.59; Eyetracking, 2011). Participants sat 71–78 cm from the eye tracker, which was placed 2 cm below the bottom edge of the monitor. Participants completed a 5-point calibration procedure before each recording and the accuracy of the calibration was verified by the experimenter before proceeding.

UAV command and control testbed and tasks

The same UAV command and control testbed from Chapter 2 was used for this chapter's research questions.

Testbed scenarios

The same testbed scenarios from Chapter 2 were used for this chapter's research questions.

Procedures

The same procedures from Chapter 2 were relevant for the present research questions. Additionally, participants completed a 5-point calibration procedure before each eye tracking recording and the accuracy of the calibration was verified by the experimenter before proceeding to the next testbed scenario.

Results

Data reduction

For calculating coefficient *K* over time, the raw gaze points, which consist of the positional (x_i, y_i) and temporal information (t_i), for 19 participants was preprocessed by custom VBA scripts, where missing and invalid data (e.g., coordinates outside the screen and blinks) was removed. The mean data loss across all participants and trials was 11.9%, (*SD*=11.2%). It was then smoothed with a second-order Butterworth filter, with a 60 Hz sampling and 6.15 Hz cutoff frequency. A velocity-threshold algorithm (I-VT; Salvucci & Goldberg, 2000) was used to distinguish saccades from fixations. Saccades were any eye movement above the velocity threshold of 20°/s; otherwise, all other data points were classified as fixations. This procedure matched the one used to empirically validate coefficient *K* (Krejtz et al., 2016).

For comparing the set of scan-based eye tracking metrics (Table 3.1) between constant and transitioning workload, the raw gaze points were screened to meet data quality requirements as outlined in ISO/TS 15007-2:2014-09, which states that at most 15% data loss is acceptable for good quality data. Following this guideline, the eye tracking of five participants was not used in any of the analyses. The mean data loss of the included participants was 7.1%. The gaze points from the eye tracker were used to calculate fixations and saccades (the eye tracker automatically filters out blinks and any fixations outside the screen were discarded). A dispersion-threshold algorithm (I-DT; Goldberg & Kotval, 1999) was applied to distinguish fixations from saccades: A cluster of gaze points was classified as a fixation if the points within the cluster were within 75 pixels of each other, and there was a minimum number of six gaze points within this fixation cluster. This made for a minimum fixation duration of approximately 100 ms. The first gaze point outside the 75-pixel limit was considered to be not part of the fixation; the gaze point just before would be the endpoint of the fixation. Any gaze points that were not part of fixations were assumed to be saccades. The calculated fixations were then used to calculate the metrics described in Table 3.1.

Analysis

All eye tracking metrics were analyzed with a RM ANOVA. To analyze how coefficient *K* compared between low and high workload periods over time for each transition rate (RQ 3.1), a 2×2×5 repeated measures analysis of variance (RM ANOVA) was used, which matches how the performance results were analyzed. Significance was set at α=0.05 and post-hoc tests with Tukey's adjustment method were used to determine significant differences between means. Violations of normality were assessed prior to analysis and Greenhouse-Geisser corrections were used when sphericity was violated. RStudio 1.2.1335 was used for all analyses (RStudio Team, 2020). General eta-squared (η_G^2) is reported as a measure of effect size for the omnibus test and values of 0.01, 0.06, and 0.14 are interpreted as small, medium, and large effect sizes, respectively (Cohen, 1988). For all post hoc pairwise comparisons, effect size was measured with Cohen's d for repeated measures (d_{rm};Lakens, 2013) and values of .2, .5, and .8 are interpreted as small, medium, and large effect sizes (Cohen, 1988). For this figure, braces indicate groupings of workload periods, while brackets with an asterisk indicate a significant comparison between the workload periods.

To compare transitioning and constant workload scenarios (RQ 3.2), a one-way RM ANOVA with four levels (i.e., the four testbed scenarios from Ch. 2) was used to compare the scan-based metrics in Table 3.1. Bonferroni corrections were applied for all post hoc tests. In all cases, Epsilon (ε) was calculated according to Greenhouse-Geisser and used to correct the one-way repeated measures ANOVA. For this set of figures, braces indicate groupings of testbed scenarios, while brackets with an asterisk indicate a significant comparison between the testbed scenarios.

RQ 3.1: Comparing coefficient K across the workload periods

To fully address RQ 3.1, i.e., the explanatory power of coefficient *K* on the observed workload history effects in Chapter 2, we examined the three-way interaction effect between transition rate, workload level, and workload period. There was a significant three-way interaction between transition rate, workload level, and period ($F(4,72)=6.568$, $p=.0001$, $\eta_G^2=0.064$). This interaction effect specifically addresses our third research question. For medium transitions, the 2nd low workload period ($M=0.15$, $SD=0.19$) had significantly higher coefficient *K* values than all other low workload periods (1st period: $M=-0.19$, $SD=0.16$, $p<.0001$, $d_{rm}=1.922$; 3rd period: $M=-0.13$, $SD=0.16$, $p<.0001$, $d_{rm}=1.547$; 4th period: $M=-0.15$, $SD=0.17$, $p<.0001$, $d_{rm}=1.587$; 5th period: $M=-0.02$, $SD=0.16$, $p=.028$, $d_{rm}=0.955$). In addition, coefficient *K* for the 2nd low workload period was also significantly higher than the 2nd low workload period of fast transitions ($M=-0.06$, $SD=0.11$, $p=.001$, $d_{rm}=1.270$). Also with medium transitions, the 1st low workload period had a significantly lower coefficient *K* value than the 5th low workload period ($p=.024$, $d_{rm}=1.06$). Figure 3.1 shows coefficient *K* values over time based on workload and transition rate.

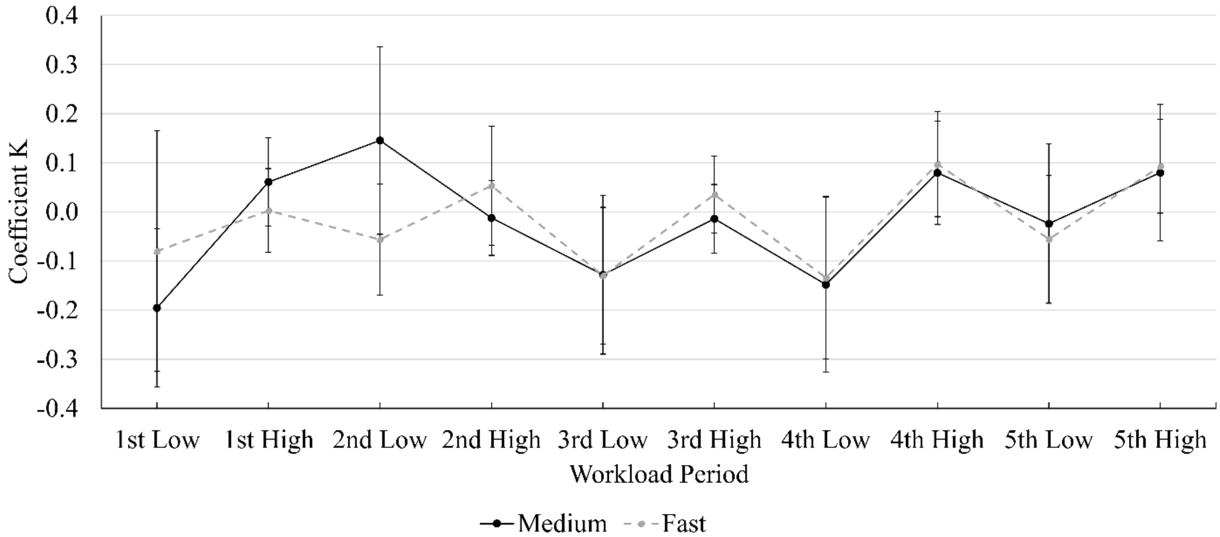


Figure 3.1 Mean coefficient *K* values for each workload period in the medium and fast transitions scenarios. Error bars represent standard deviation of the mean

RQ 3.2: Comparing scan-based metrics to constant workload

The spread metrics for each testbed scenario are presented in Figure 3.2. There was no significant effect of testbed scenario on convex hull area (Figure 3.2a). There was a significant effect of testbed scenario on spatial density (F(2.48,37.24)=5.24, p=.006, η_p^2=.25, ε=.82; Figure 3.2b). Post hoc tests found the low workload scenario had a larger spatial density than the medium (p=.004) and fast transition scenario (p=.008). There was also a significant effect of testbed scenario on stationary gaze entropy (F(2.21,33.17)=9.49, p<.001, η_p^2=.38, ε=.78; Figure 3.2c). Post hoc tests found that low workload had a higher stationary gaze entropy than all other testbed scenarios (high workload: p=.009, medium transition: p=.001, and fast transition: p=.007). Stationary gaze entropy was also significantly higher in the high workload scenario than the medium transition scenario (p=.014).

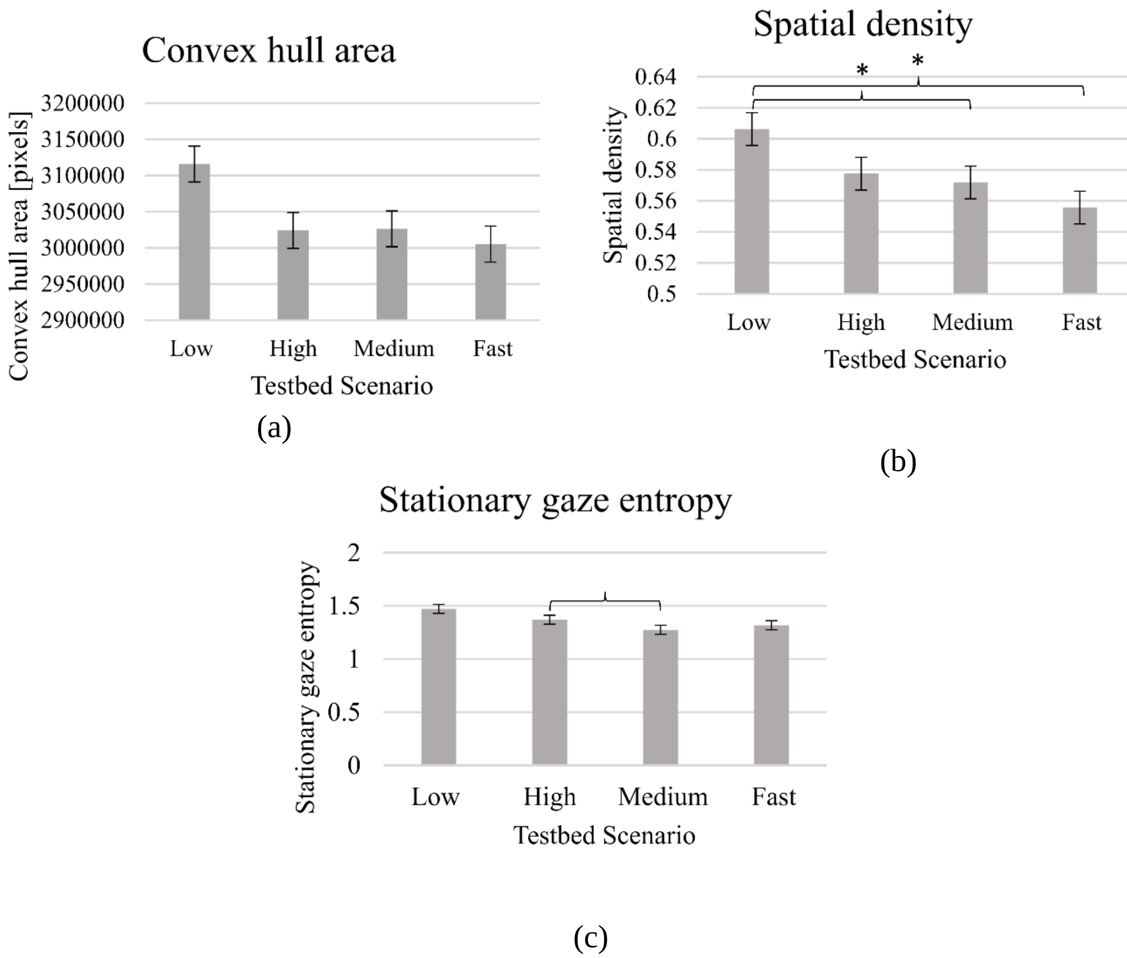


Figure 3.2 Results of the spread metrics for each testbed scenario: (a) convex hull area, (b) spatial density, and (c) stationary gaze entropy. Error bars represent standard error of the mean

For the directness metrics, which can be seen in Figure 3.3, there was a significant effect of testbed scenario on mean saccade amplitude (F(2.38,35.76)=13.81, p<.001, η_p^2=.47, ε=.79; Figure 3.3a). Post hoc tests showed that the low workload scenario had significantly larger saccade amplitude than all other testbed scenarios (all p<.001). For scanpath length per second, there was also a significant effect of testbed scenario (F(2.84,42.61)=16.32, p<.001; η_p^2=.52, ε=.94; Figure 3.3b). Post hoc tests showed that there was a significant difference between low workload scenario and all other testbed scenarios (all $p<.001$). There was no significant effect of testbed scenario on backtrack rate (Figure 3.3c). There was a significant effect of testbed scenario on gaze transition rate ($F(2.65,39.75)=4.92$, $p=.007$, $\eta_p^2=.24$, $\varepsilon=.88$; Figure 3.3d). Post hoc tests found gaze transition rate was significantly lower during the low workload scenario compared to the medium ($p=.029$) and fast transitions scenario ($p=.005$). For gaze transition entropy, there was an effect of testbed scenario ($F(2.16,32.48)=15.83$, $p<.001$, $\eta_p^2=.51$, $\varepsilon=.72$; Figure 3.3e). Gaze transition entropy was significantly higher in the low workload scenario than all other testbed scenarios (all $p=<.001$). The gaze transition entropy in the high workload scenario was also significantly higher than the medium transition scenario ($p=.007$). Finally, for the duration metric, there was no significant effect of testbed scenario on mean fixation duration (Figure 3.4).

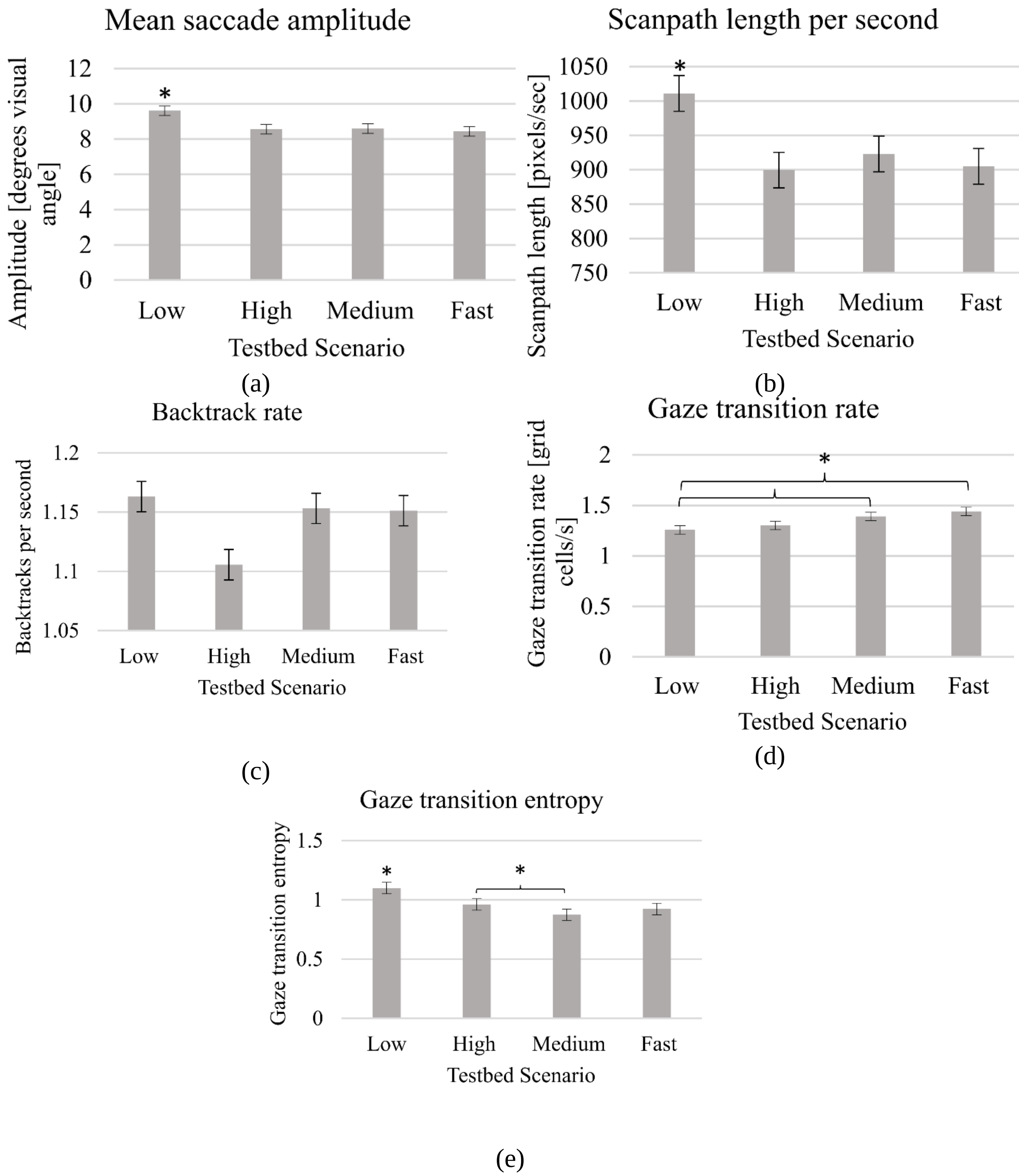


Figure 3.3 Results of the directness metrics for each testbed scenario: (a) mean saccade amplitude, (b) scanpath length per second, (c) backtrack rate, (d) gaze transition rate, and (e) gaze transition entropy. Error bars represent standard error of the mean

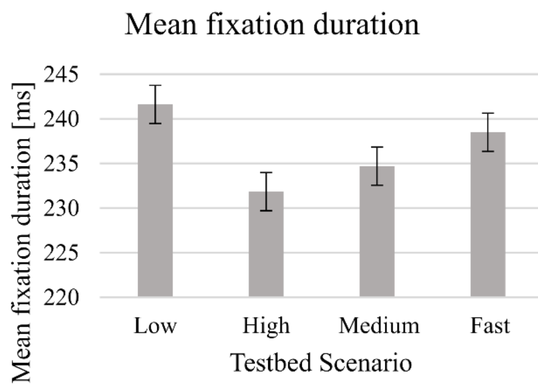


Figure 3.4 Results of the duration metric for each testbed scenario. Error bars represent standard error of the mean

Table 3.2 summarizes the present eye tracking results across workload transitions and constant workload.

Table 3.2 Summary of the eye tracking analysis between transitioning and constant workload

Metric	**Result**	**Main conclusions**
	Spread Metrics	
Convex hull area (pixels^2)	Not significant	Increased spread was associated with worst performance
Spatial density	• Highest value in low workload compared to medium and fast transitions	
Stationary gaze entropy	• Highest value in low workload • Higher value in high workload than medium transitions	
	Directness Metrics	
Mean saccade amplitude (pixels)	• Highest value in low workload scenario	Except for gaze transition rate, all of metrics indicated less directness (i.e., efficiency) was associated with worst performance
Scanpath length per second (pixels)	• Highest value in low workload scenario	
Backtrack rate [/sec]	Not significant	
Gaze transition rate [grid cells/sec]	• Lowest value in low workload scenario compared to medium and fast transitions	
Gaze transition entropy	• Highest value in low workload • Higher value in high workload than medium transitions	
	Duration Metric	
Fixation duration [ms]	Not significant	Duration of visual attention was not associated with any performance trend

Discussion

The aim of this chapter was to understand if and how visual attention allocation explained the workload transition performance trends. Overall, including eye tracking data, specifically scan-based metrics, was informative on the performance trends observed in Chapter 2 and specifics as to how will be discussed by addressing the outcomes of each research question.

RQ 3.1: Is coefficient *K* informative of the workload history effects observed during medium and fast transitions?

Most of the performance differences between medium and fast transitions were based on workload period, i.e., their respective workload history effect (Ch. 2). To better understand why this was the case, we consider the performance and coefficient *K* results together, to see whether these two data streams can provide a more complete understanding of the workload history effect. Overall, our expectations for this research question were partially met when considering the result of fast transitions: there was a decline in performance when participants engaged in more ambient visual attention initially, but performance improved later after coefficient *K* became aligned with workload level expectations, i.e., positive coefficient *K* during high workload periods and negative coefficient *K* during low workload periods. The findings for fast transitions show that a participant's performance can recover if visual attention allocation strategies are developed in accordance with each workload level.

Oppositely, for medium transitions, coefficient *K* peaked during the 2nd low workload period as it was significantly higher than all other low workload periods and the corresponding low workload period of fast transitions. The significant increase in coefficient *K* coincides with the first time low workload periods of medium transitions irreversibly increase in response times and decrease in accuracy rates. Workload history effects with medium transitions may be due to a large, unexpected, increase in focal attention during low workload. Similarly, coefficient *K* was significantly larger in the final low workload period of medium transitions compared to its first low workload period. The findings here may indicate that an increase in focal attention during low workload periods of workload transitions may be an indicator of future performance decrements, i.e., a workload history effect, and not an improved acquaintance with the environment like previous work suggests (Irwin & Zelinsky, 2002; Krejtz et al., 2017; Krejtz et al., 2018). Workload history effects are more likely to be apparent during low workload periods (Bowers et al., 2014; Cox-Fuenzalida, 2007; Matthews, 1986), but now coefficient *K* provides a quantitative explanation as to why this is the case—not using the appropriate visual attention type as a function of workload. Furthermore, even though coefficient *K* values eventually converged to similar patterns over time for both transition rates, performance trends did not. One potential explanation could be adopting ineffective strategies at the onset may have both immediate and delayed effects on performance, even if more effective strategies are adopted eventually, as presently seen with medium transitions. It may also show that the visual attention type adopted for fast transitions may not be best for medium transitions and vice versa. A follow-up study should ideally examine workload history over a longer time frame to see if performance recovers when an effective visual attention management strategy is eventually adopted and if effective visual attention allocation varies by transition rate.

Overall, these results suggest that it is important to immediately adopt an effective strategy to account for workload transitions or performance may irrevocably suffer (i.e., a workload history effect endures). The findings also support the premise that people initially evaluate the environment to develop a management strategy for workload transition, but struggle to update this strategy over time (Over et al., 2007; Prytz & Scerbo, 2015). The findings add to the effort regulation explanation because how operators manage workload transitions impacts performance over the course of workload transitions. Previous research that examined workload history with other psychophysiological measures detect a cognitive response to the workload transition, but the practical interpretation of these measures is unclear (Bowers et al., 2014; Kim et al., 2019). However, the findings here show the potential of advanced scan-based metrics to be used in real-time and inform the design of intelligent technology (Feigh et al., 2012; Rothrock et al., 2002).

RQ 3.2: Are scan-based metrics informative on the performance differences between transitioning and constant workload?

In Chapter 2, the analysis of the primary and secondary tasks performance during both low and high workload provided further insights into the effects of workload transitions. There were minimal performance differences between primary and secondary task performance of medium and fast transitions when analyzed overall and by workload level. Similarly, none of the scan-based metrics were significantly different between medium and fast transitions. It appears the small subset of overall and by workload level performance differences between the two transition scenarios is not a function of how participants sampled and extracted information from the environment. As for how it compares to constant workload, primary and secondary tasks mostly performed better during workload transitions. The results confirm that workload transitions can affect how people perform their tasks in a way that is different from just high or low workload (Cox-Fuenzalida, 2007; Cumming & Croft, 1973; Goldberg & Stewart, 1980), especially when in a dynamic, realistic, multitasking environment (Jansen et al., 2016). Analyzing the scan-based measures aims to understand how priorities and strategies adjusted during workload transitions to prompt superior multitasking performance.

In general, results were consistent with the hypothesis of worst multitasking performance being associated with increased spread and less directness. As previously discussed in Chapter 2, worse multitasking performance was surprisingly during the low and high workload condition. For example, during low workload, the spread metrics suggest participants were covering wider and more varied areas of the display, whereas the directness metrics suggest participants were scanning less efficiently especially compared to medium and fast transitions. The only exception to this statement is gaze transition rate being higher with medium and fast transitions than compared to constant low workload. Similarly, the only expectation that was not met was for the duration metric; results suggest that there was no association between certain performance findings or workload transitions and the length of time participants spent fixating. This is surprising given how frequently mean fixation duration is used in the context of workload (e.g., Schulz et al., 2011). It could be that there is no issue of discriminating information in this study, although this would need to be further explored.

Multitasking performance trends were best reflected in the eye tracking metrics capturing transitions between AOIs. Specifically, stationary gaze entropy and gaze transition entropy were the only measures where both the constant low and high workload scenarios were significantly different than the transitioning scenarios, suggesting that these metrics are particularly sensitive to changes occurring within the environment. These two metrics are based on the Markov property, a stochastic process in which the next state (in this case, the next AOI that is fixated) depends only on the current state (current AOI being fixated). By modeling visual attention allocation as probabilities of *where* attention will switch based on the current location of attention, it directly informs task strategy, which seems to be the key difference between constant and transitioning workload performance. These measures found better performance was associated with more concentrated and routine attention transitions, respectively. It would appear that the order attention transitions to AOIs and the probability of it being on each AOI is an aspect that should be regularly explored in studies where workload transitions in dynamic environments, contrary to what is currently the case.

Visual attention allocation was more structured and planned during both workload transition rates, potentially suggesting workload transitions assisted with the regulation of mental resources. The spread metrics suggest that visual attention was covering more of the display during the low workload scenario (e.g., higher spatial density) and the distribution of transitions across panels was larger (e.g., higher stationary gaze entropy) for both the low and high workload scenarios compared to the medium and fast transitions scenarios. The directness metrics suggested participants were scanning less in general (e.g., decreased gaze transition rate) and with less purpose (e.g., higher gaze transition entropy) in the low and/or high workload scenario compared to the medium and fast transitions scenarios. Therefore, the strategic use of mental resources contributed to the improved performance during workload transitions which supports the effort regulation explanation because improved performance is associated with the ability to strategically deploy mental resources, i.e., establish a visual attention allocation strategy. A more systematic visual attention allocation strategy used during workload transitions would also explain why there was an increase in gaze transition rate: participants moved from one general area of the screen to another in an urgent, systematic fashion. During low workload, participants scanned more of the display with less purpose, potentially making their task completion strategy less systematic. This seemed to be a bigger contributor in hindering effective multitasking performance rather than the amount of mental resource available for use, which challenge the applicability of the resource depletion explanation. The findings here demonstrate the importance of considering multiple types of scan-based metrics to understand the effects of workload transitions on visual attention allocation as thoroughly as possible.

Overall, Chapter 3 shows how scan-based eye tracking measures help explain the operator's process in managing workload transitions and its impact on performance. Specifically, the spread, directness, and type of visual attention can be a non-invasive, psychophysiological, and quantitative method to further understand *how* people are impacted by and responding to workload transitions, and how this leads to certain workload history effects and multitasking performance trends.

Limitations

More work is needed before concretely relating scan-based measures to workload transition performance trends. First, a study with a similar task paradigm, but larger sample size and different display layout should be completed to see if these results replicate. Second, it might be best to investigate how different lengths of time at each workload level impacts coefficient *K* results. Here, medium and fast transitions had different amounts of time in low workload due to the experimental setup and it is unclear if this influenced the coefficient *K* results (Krejtz et al., 2016). Two potential ways it could potentially impact the comparisons made between the two transition rates include (a) coefficient *K* values in the low workload periods of the medium transition were more sensitive to any major fluctuations in fixation duration and/or saccade amplitude or (b) the longer durations in the low workload periods of fast transitions gave participants more time to establish a consistent visual attention strategy for each low workload period. Albeit, this is a limitation of the present study, coefficient *K* still appears to be a promising metric in capturing changes to visual attention and performance outcomes over time when workload changes.

It would also be interesting to look at how the set of scan-based metrics in Table 3.1 detail performance trends over time. This would be particularly important when studying the effects of different workload transition rates given this is where most of those performance differences manifested. Exploring over longer workload periods would also better assess the applicability of the resource depletion explanation, as more time may be needed for resource to deplete and/or recover (Gluckman et al., 1993). However, most previous workload transition research is conducted in a time frame that is similar to the present work and both theories have been found to apply (e.g., Cox-Fuenzalida & Angie, 2005; Hancock et al., 1995; Jansen et al., 2016; Morgan & Hancock, 2011). Furthermore, the sampling rate of the eye tracker used in this study (60 Hz) is not ideal for the study of saccade amplitude, although it makes no difference to the detection of fixations or the coefficient *K* calculation (Leube et al., 2017; Krejtz et al., 2016). Finally, future work should thoroughly examine the explanatory power of scan-based metrics by modeling them as predictors of workload transition performance. Conducting this analysis will not only verify the currently observed associations between visual attention allocation patterns and performance, but it will be necessary if eye tracking is to serve as the basis of advanced technology that can cater to the needs of the individual in real-time.

Conclusion

Coefficient *K* has promise to be a real-time, proactive indicator of visual attention strategies that might later lead to negative workload history effects during low workload. Therefore, coefficient *K* may be a reliable indicator of when the operator needs assistance and how to provide that assistance (e.g., encourage the operator to engage in a certain visual attention type when workload transitions over time). This potential has not been as promising with other real-time, cognitive-based measures (e.g., EEG; Bowers et al., 2014; Kim et al., 2019). Additionally, using a combination of scan-based metrics, especially ones capturing the context of the environment, assists in explaining how visual attention allocation impacts performance during workload transitions. In particular, the spread (namely, spatial density and stationary gaze entropy) and directness metrics (namely, gaze transition rate and gaze transition entropy) provided an explanation on why workload transitions outperformed constant workload, i.e., visual attention was more concentrated and efficient. This is interesting considering these types of metrics are rarely used in the context of workload (e.g., Coyne et al., 2017; Foy & Chapman, 2018). The two entropy metrics were particularly informative, suggesting that they should be used more in studies of workload transitions. Given that stationary gaze entropy reflects where visual attention was most likely to be across the different tasks, these metrics provide direct, quantifiable insight into participants' task switching behavior during workload transitions – a completely novel finding. It would seem workload transitions cannot be expected to produce performance trends that are some "average" between low and high workload, as workload transitions seem to prompt a different multitasking approach.

Our future work will expand upon these findings for the benefit of both theory and application. For example, if workload transition performance is a function of how visual attention is deployed, then the effort regulation explanation should be further expanded upon. Additionally, understanding how visual attention impacts performance, directly informs the design of visual displays in the environment, and how to further investigate the design of advanced technology, like adaptive displays (Feigh et al., 2012).

Next Steps

Chapters 2 and 3 served as initial investigations into the effects of workload transition rate on performance and visual attention allocation patterns. The biggest findings so far include:

- Workload transitions in multitasking, realistic environments produce nuanced, unexpected performance trends potentially because they prompt the active regulation of mental resources.
- Transition rate differences were the most prominent with time-based analyses.
- Advanced scan-based metrics were the most informative on task strategy and performance differences over time.

Per these results and their identified limitations, our follow up investigation includes:

- A further expansion on workload transition rate with a context-relevant population.
- Analyzing workload transition performance with longitudinal analysis methods, while actively accounting for the performance trends of each individual.
- Directly synthesizing the performance and visual attention allocation together by making the latter a direct predictor in the former's trends over time.

The follow up study and analysis fills out the remaining two chapters of this dissertation.

References

Bowers, M. A., Christensen, J. C., & Eggemeier, F. T. (2014). The effects of workload transitions in a multitasking environment. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 220–224. https://doi.org/10.1177/1541931214581046

Buswell, G. T. (1935). How people look at pictures: A study of the psychology and perception in art. Univ. Chicago Press.

Cain, B. (2007). *A review of the mental workload literature*. Defence Research and Development Canada Toronto.

Cohen, J. (1988). The t test for means. In *Statistical power analysis for the behavioral sciences* (pp. 19–74). Lawrence Erlbaum Associates, Inc., Publishers.

Coral, M. P. (2016). *Analyzing cognitive workload through eye-related measurements: A meta-analysis*. [Master's thesis, Wright State University]. CORE Scholar.

Cox-Fuenzalida, L.-E. (2007). Effect of workload history on task performance. *Human Factors 49*(2), 277–291.

Cox-Fuenzalida, L.-E., Angie, A., Holloway, S., & Sohl, L. (2006). Extraversion and task performance: A fresh look through the workload history lens. *Journal of Research in Personality*, *40*(4), 432–439. https://doi.org/10.1016/j.jrp.2005.02.003

Cox-Fuenzalida, L.-E., Swickert, R., & Hittner, J. B. (2004). Effects of neuroticism and workload history on performance. *Personality and Individual Differences*, *36*(2), 447–456. https://doi.org/10.1016/S0191-8869(03)00108-9

Coyne, J.T., Sibley, C., Sherwood, S., Foroughi, C.K., Olson, T., & Vorm, E. (2017). Assessing workload with low cost eye tracking during a supervisory control task. *International*

Conference on Augmented Cognition, 139–147. https://doi.org/10.1007/978-3-319-58628-1_12

Cumming, R. W., & Croft, P. G. (1973). Human information processing under varying task demand. *Ergonomics*, *16*(5), 581–586. https://doi.org/10.1080/00140137308924548

Cutrell, E., & Guan, Z. (2007). What are you looking for? An eye-tracking study of information usage in Web search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*, 407–417. https://doi.org/10.1145/1240624.1240690

De Rivecourt, M., Kuperus, M. N., Post, W. J., & Mulder, L. J. M. (2008). Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics*, *51*(9), 1295–1319. https://doi.org/10.1080/00140130802120267

Devlin, S. P., Byham, J. K., & Riggs, S. L. (2021). Does what we see shape history? Examining workload history as a function of performance and ambient/focal visual attention. *Transactions on Applied Perception*, *8*(2), 1-17. https://doi.org/10.1145/3449066

Di Nocera, F., Camilli, M., & Terenzi, M. (2007). A random glance at the flight deck: Pilots' scanning strategies and the real-time assessment of mental workload. *Journal of Cognitive Engineering and Decision Making*, *1*(3), 271–285. https://doi.org/10.1518%2F155534307X255627

Di Stasi, L. L., Antolí, A., & Cañas, J. J. (2013). Evaluating mental workload while interacting with computer-generated artificial environments. *Entertainment Computing*, *4*(1), 63–69. https://doi.org/10.1007/s00464-016-4851-8

Di Stasi, L. L., Diaz-Piedra, C., Rieiro, H., Sánchez Carrión, J. M., Martin Berrido, M., Olivares, G., & Catena, A. (2016). Gaze entropy reflects surgical task load. *Surgical Endoscopy*, *30*(11), 5034–5043. https://doi.org/10.1007/s00464-016-4851-8

Duchowski, A. T., Krejtz, K., Gehrer, N. A., Bafna, T., & Bækgaard, P. (2020). The low/high index of pupillary activity. *CHI '20*, 1–12. https://doi.org/10.1145/3313831.3376394

Edwards, T., Gabets, C., Mercer, J., & Biener, N. (2017). Task demand variation in air traffic control: Implications for workload, fatigue, and performance. In *Advances in Human Aspects of Transportation* (pp. 91–102). Springer. https://doi.org/10.1007/978-3-319-41682-3

Estes, S. (2015). The workload curve: Subjective mental workload. *Human Factors*, *57*(7), 1174–1187. https://doi.org/10.1177/0018720815592752

Eyetracking. (2011). Hardware: Eye tracking systems. http://www.eyetracking.com/Hardware/Eye- Tracker- List

Fallahi, M., Motamedzade, M., Heidarimoghadam, R., Soltanian, A. R., Farhadian, M., & Miyake, S. (2016). Analysis of the mental workload of city traffic control operators while monitoring traffic density: A field study. *International Journal of Industrial Ergonomics*, *54*, 170–177. https://doi.org/10.1016/j.ergon.2016.06.005

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. https://doi.org/10.1177/0018720812443983

Foy, H. J., & Chapman, P. (2018). Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation. *Applied Ergonomics*, *73*, 90–99.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review*, *23*(4), 523–552. https://doi.org/10.1007/s10648-011-9174-7

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: Methods and constructs. *International Journal of Industrial Ergonomics*, *24*(6), 631–645.

https://doi.org/10.1016/S0169-8141(98)00068-7

Goldberg, R. A., & Stewart, M. R. (1980). Memory overload or expectancy effect? 'Hysteresis' revisited. *Ergonomics*, *23*(12), 1173–1178. https://doi.org/10.1080/00140138008924824

Hampson, R. E., Opris, I., & Deadwyler, S. A. (2010). Neural correlates of fast pupil dilation in nonhuman primates: Relation to behavioral performance and cognitive workload. *Behavioural Brain Research*, *212*(1), 1–11. https://doi.org/10.1016/j.bbr.2010.03.011

Hancock, P. A., Williams, G., & Manning, C. M. (1995). Influence of task demand characteristics on workload and performance. *The International Journal of Aviation Psychology*, *5*(1), 63–86. https://doi.org/10.1207/s15327108ijap0501_5

Helton, W. S., Shaw, T., Warm, J. S., Matthews, G., & Hancock, P. (2008). Effects of warned and unwarned demand transitions on vigilance performance and stress. *Anxiety, Stress, and Coping*, *21*(2), 173–184. https://doi.org/10.1080/10615800801911305

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive energetical framework. *Biological Psychology*, *45*, 73–93. https://doi.org/10.1016/S0301-0511(96)05223-4

Hwang, S. L., Yau, Y. J., Lin, Y. T., Chen, J. H., Huang, T. H., Yenn, T. C., & Hsu, C. C. (2008). Predicting work performance in nuclear power plants. *Safety Science*, *46*(7), 1115–1124. https://doi.org/10.1016/j.ssci.2007.06.005

Irwin, D. E., & Zelinsky, G. J. (2002). Eye movements and scene perception: Memory for things observed. *Perception and Psychophysics*, *64*(6), 882–895. https://doi.org/10.3758/BF03196793

Jacob, R. J. K., & Karn, K. S. (2003). Eye tracking in human–computer interaction and usability research: Ready to deliver the promises. In *In The Mind's Eye: Cognitive and Applied*

Aspects of Eye Movement Research (pp. 573–607).

Jansen, R. J., Sawyer, B. D., van Egmond, R., de Ridder, H., & Hancock, P. A. (2016). Hysteresis in mental workload and task performance: The influence of demand transitions and task prioritization. *Human Factors*, *58*(8), 1143–1157. https://doi.org/10.1177/0018720816669271

Jarodzka, H., Scheiter, K., Gerjets, P., & van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learning and Instruction*, *20*(2), 146–154. https://doi.org/10.1016/j.learninstruc.2009.02.019

Jiang, J., He, D., & Allan, J. (2014). Searching, browsing, and clicking in a search session. *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 607–616. https://doi.org/10.1145/2600428.2609633

Kim, N., House, R., Yun, M. H., & Nam, C. S. (2019). Neural correlates of workload transition in Multitasking: An ACT-R model of hysteresis effect. *Frontiers in Human Neuroscience, 12*(535), 1–14. https://doi.org/10.3389/fnhum.2018.00535

Krejtz, K., Çöltekin, A., Duchowski, A. T., & Niedzielska, A. (2017). Using coefficient *K* to distinguish ambient/focal visual attention during cartographic tasks. *Journal of Eye Movement Research*, *10*(2). https://doi.org/10.16910/jemr.10.2.3

Krejtz, K., Duchowski, A., Krejtz, I., Szarkowska, A., & Kopacz, A. (2016). Discerning ambient/focal attention with coefficient *K*. *ACM Transactions on Applied Perception*, *13*(3), 1–20. https://doi.org/10.1145/2896452

Krejtz, K., Duchowski, A., Szmidt, T., Krejtz, I., González Perilli, F., Pires, A., Vilaro, A., & Villalobos, N. (2015). Gaze transition entropy. *ACM Transactions on Applied Perception*, *13*(1), 1–20. https://doi.org/10.1145/2834121

Krejtz, K., Wisiecka, K., Krejtz, I., Holas, P., Olszanowski, M., & Duchowski, A. T. (2018). Dynamics of emotional facial expression recognition in individuals with social anxiety. *Eye Tracking Research & Applications Symposium (ETRA)*, 43, 1-9. https://doi.org/10.1145/3204493.3204533

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(863), 1–12. https://doi.org/10.3389/fpsyg.2013.00863

Leube, A., Rifai, K., & Wahl, S. (2017). Sampling rate influences saccade detection in mobile eye tracking of a reading task. *Journal Eye Movement Research, 10*(3), 4–14. https://doi.org/10.16910/jemr.10.3.3

Matthews, M. L. (1986). The influence of visual workload history on visual performance. *Human Factors*, *28*(6), 623–632. https://doi.org/10.1177%2F001872088602800601

Matthews, G., De Winter, J., & Hancock, P. A. (2019). What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. *Theoretical Issues in Ergonomics Science*, *21*(4), 369–396. https://doi.org/10.1080/1463922X.2018.1547459

McKendrick, R. D., & Cherry, E. (2018). A deeper look at the NASA TLX and where it falls short. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 44–48. https://doi.org/10.1177/1541931218621010

Moacdieh, N. M., Devlin, S. P., Jundi, H., & Riggs, S. L. (2020). Effects of workload and workload transitions on attention allocation in a dual-task environment: Evidence from eye tracking metrics. *Journal of Cognitive Engineering and Decision Making*, *14*(2), 132–151. https://doi.org/10.1177/1555343419892184

Moacdieh, N., & Sarter, N. (2015). Display clutter: A review of definitions and measurement techniques. *Human Factors*, *57*(1), 61–100. https://doi.org/10.1177%2F0018720814541145

Monfort, S. S., Sibley, C. M., & Coyne, J. T. (2016). Using machine learning and real-time workload assessment in high-fidelity UAV simulation environment. *Proceedings of SPIE*, https://doi.org/10.1117/12.2219703

Morgan, J. F., & Hancock, P. A. (2011). The effect of prior task loading on mental workload: An example of hysteresis in driving. *Human Factors, 53*(1), 75–86. https://doi.org/10.1177/0018720810393505

Over, E. A. B., Hooge, I. T. C., Vlaskamp, B. N. S., & Erkelens, C. J. (2007). Coarse-to-fine eye movement strategy in visual search. *Vision Research*, *47*(17), 2272–2280. https://doi.org/10.1016/j.visres.2007.05.002

Pan, B., Hembrooke, H. A., Gay, G. K., Granka, L. A., Feusner, M. K., & Newman, J. K. (2004). The determinants of web page viewing behavior. *Proceedings of the Eye Tracking Research & Applications Symposium*, 147–154. https://doi.org/10.1145/968363.968391

Pannasch, S., Helmert, J. R., Roth, K., & Walter, H. (2008). Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions. *Journal of Eye Movement Research*, *2*(2), 1–19. https://doi.org/10.16910/jemr.2.2.4

Poole, A., & Ball, L. J. (2006). Eye tracking in human-computer interaction and usability research. In A. Poole & L.J. Ball (Eds.), *Encyclopedia of human computer interaction* (pp. 211–219). Idea Group.

Previc, F. H. (1998). The neuropsychology of 3-D space. *Psychological Bulletin*, *124*(2), 123–164. https://doi.org/10.1037/0033-2909.124.2.123

Prytz, E. G., & Scerbo, M. W. (2015). Changes in stress and subjective task load over time following a task load transition. *Theoretical Issues in Ergonomics Science*, *16*(6), 586–605. https://doi.org/10.1080/1463922X.2015.1084397

Rantanen, E. M., & Goldberg, J. H. (1999). The effect of mental task load on the visual field size and shape. *Ergonomics*, *42*(6), 816–834. https://doi.org/10.1080/001401399185315

Rothrock, L., Koubek, R., Fuchs, F., & Haas, M. (2002). Reappraisal of adaptive interfaces: Toward biologically inspired paradigms. *Theoretical Issues in Ergonomics Science*, *3*(1), 47–84. https://doi.org/10.1080/14639220110110342

Rozado, D., & Dunser, A. (2015). Combining EEG with pupillometry to improve cognitive workload detection. Computer, *48*(10), 18–25. https://doi.org/10.1109/MC.2015.314

RStudio Team (2020). *RStudio: Integrated Development for R.* RStudio, PBC, Boston, MA http://www.rstudio.com/

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the Eye Tracking Research & Applications Symposium*, 71–78. https://doi.org/10.1145/355017.355028

Savage, S. W., Potter, D. D., & Tatler, B. W. (2013). Does preoccupation impair hazard perception? A simultaneous EEG and eye tracking study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *17*, 52–62. https://doi.org/10.1016/j.trf.2012.10.002

Schulz, C. M., Schneider, E., Fritz, L., Vockeroth, J., Hapfelmeier, A., Wasmaier, M., Kochs, E. F., & Schneider, G. (2011). Eye tracking for assessment of workload: A pilot study in an anaesthesia simulator environment. *British Journal of Anaesthesia*, *106*(1), 44–50. https://doi.org/10.1093/bja/aeq307

Shiferaw, B., Downey, L., & Crewther, D. (2019). A review of gaze entropy as a measure of visual scanning efficiency. *Neuroscience and Biobehavioral Reviews*, *96*, 353–366.

https://doi.org/10.1016/j.neubiorev.2018.12.007

Shiferaw, B., Downey, L. A., Westlake, J., Stevens, B., Rajaratnam, S. M. W., Berlowitz, D. J., Swann, P., & Howard, M. E. (2018). Stationary gaze entropy predicts lane departure events in sleep-deprived drivers. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-20588-7

Smeets, J. B. J., & Hooge, I. T. C. (2003). Nature of variability in saccades. *Journal of Neurophysiology*, *90*(1), 12–20. https://doi.org/10.1152/jn.01075.2002

Velichkovsky, B. M., Joos, M., Helmert, J. R., & Pannasch, S. (2005). Two visual systems and their eye movements: Evidence from static and dynamic scene perception. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2283–2288.

Veltman, J. A., & Gaillard, A. W. K. (1996). Physiological indices of workload in a simulated flight task. *Biological Psychology*, *42*(3), 323–342. https://doi.org/10.1016/0301-0511(95)05165-1

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: Mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17. https://doi.org/10.1080/00140139.2014.956151

CHAPTER 4

How does workload transition rate influence performance trends over time?

Introduction

Per the performance trends in Chapter 2, this chapter aims to understand how performance trends *over time*. It specifically investigates if these performance trends depend on the transition rate, the workload level, i.e., low versus high workload, and/or the individual managing the workload transition. Two different analysis methods, one that is widely relied upon and one that is new to workload transition research, will be conducted to provide practical, nuanced, and operator-centric design guidance. Chapter 4 begins to simultaneously address the applied research gaps stated at the end of Chapter 3 as a means to develop theory and design, given it has been limited to sporadic and vague recommendations on the expected outcomes of workload transitions, with no guidance on how operators should manage workload transitions in real-world environments. Specifically, Chapter 4 details how primary task performance trends over time across three different transition rates.

Background on the experimental paradigm and analysis method

Further examining the impact of transition rate and multiple transitions can potentially inform the applicability of each theoretical explanation (e.g., Ungar et al., 2005). Including these characteristics may also lead to a more thorough understanding of workload transition performance, such as speed/accuracy tradeoffs, low versus high workload performance, and the severity of decrements and/or improvements over time. Fleshing out these details can lead to more informative design guidance for environments experiencing workload transitions. We decided to further explore transition rate given workload transitions produced unexpectedly better performance and more purposeful visual attention allocation patterns.

Second, context-relevant populations, i.e., non-university student populations, are rarely recruited for workload transition research, even though there is evidence of workload transitions impacting individuals differently (Cox-Fuenzalida, Angie, et al., 2006; Cox-Fuenzalida et al., 2004; Devlin & Riggs, 2018; McKendrick & Harwood, 2019; Mracek et al., 2014). Edwards et al. (2017) studied workload transition performance of air traffic controllers and performance surprisingly improved between the two high workload periods in a high-low-high workload paradigm (Figure 1.1d). However, their sample size only included eight participants, which questions the generalizability of their results. The workload history effects found in Chapter 2 did report rather large performance variance across the workload periods, making it possible that performance trends over time may have differed across participants, but there was no way to confirm this with the analysis method used, i.e., post hoc tests of a repeated measures analysis of variance.

Third, the analysis method to date has consisted of either pairwise comparisons to a constant workload baseline (e.g., comparing performance during the low workload period in a workload transition to performance during constant low workload; Bowers et al., 2014; Cox-Fuenzalida, 2007; Devlin et al., 2020) or between the same workload levels over time (e.g., comparing 1st and 2nd low workload periods, etc.; Devlin et al., 2021; Jansen et al., 2016; Kim et al., 2019; Morgan & Hancock, 2011). Although the latter is sufficient to determine workload history effects, it cannot decipher if performance *fluctuates (temporary change)* or *changes (indefinite change)* over time and if either depends on the *individual*. Determining whether these types of trends exist could inform the theory and design surrounding workload transitions (Cox-Fuenzalida et al., 2004; Cox-Fuenzalida et al., 2006; McKendrick & Harwood, 2019; Mracek et al., 2014).

Motivation for the current chapter

The goal of Chapter 4 was to understand how transition rate, workload level, and the individual impact performance trends over time for both response time and accuracy when there are multiple instances of workload transitions. The specific research questions aim to build upon the current theoretical explanations and provide design guidance to better support operators experiencing workload transitions in complex environments. There is growing evidence workload transition performance trends are transient over time (Jansen et al., 2016) and depend on the individual (McKendrick & Harwood, 2019; Mracek et al., 2014), yet the analysis used by most existing studies cannot detect these critical caveats. One way to simultaneously address these two gaps consists of analyzing performance trends from a context-relevant population with methods such as *growth curve modeling.* This modeling approach estimates performance trends over time by estimating and comparing how each individual performs over time (Curran, Obeidat, & Losardo, 2011). Completing both the aggregate analysis (traditional analysis used in existing literature) and growth curve modeling are necessary to answer the following research questions:

1. **RQ 4.1:** Does transition rate affect performance trends over time?

a. **Expectations:** Previous work showed performance trends depended on transition rate, with slower transition rates showing performance declines over time (Devlin et al., 2021; Moroney et al., 1995). Our previous work also found performance trends depended on workload level, i.e., the low and high workload periods within a workload transition, and the performance measure, i.e., response time and accuracy. Hence, we expect response times to slow, especially across the low workload periods of each transition rate, but ultimately recover for all transition rates, where recovery will be faster and more pronounced with faster transition rates (Devlin et al., 2021; Moroney et al., 1995). For accuracy, we expect an initial decline during low and high workload periods of slower transitions, but then a slight recovery; yet with faster transitions, accuracy will remain unchanged over time (Devlin et al., 2021; Moroney et al., 1995).

2. **RQ 4.2:** Which theoretical explanation—i.e., resource depletion or effort regulation—accounts for the observed performance trends over time?

    a. **Expectations:** Given the aggregate analysis relies on pairwise comparisons, it is not equipped to indicate the general performance trends over time (i.e., deciphering fluctuation from change). Here, growth curve modeling serves as a way to concisely and uniquely quantify performance trends in that it estimates a single performance *trajectory* for each individual across both low and high workload. If resource depletion better explains the performance trends over time, then growth curve modeling will estimate a quadratic trajectory, i.e., a U-shape, for both response time and accuracy. Specifically, it will show performance initially suffering, but then recovering for all individuals. If effort regulation better explains the performance trends over time, then growth curve modeling will estimate both performance metrics to have linear trajectories, with its direction and steepness dependent on the individual.

3. **RQ 4.3:** Does the individual affect performance trends over time?

    a. **Expectations:** Chapter 2 found large performance variability in low and high workload periods for both transition rates (Devlin et al., 2021). However, our previous analysis method, i.e., the aggregate analysis, could not reveal whether performance trends varied across individuals. For this research question, we rely on growth curve modeling because it estimates a performance trajectory for each individual and then assesses the variability amongst the trajectories. Although our study is with a homogenous, context-relevant population, we expect that growth curve modeling will estimate significant variability between individual trajectories, especially for faster transition rates, given transition rate is not often studied in previous work and performance outcomes of previous work diverges (see review in Bowers, 2013). A counterintuitive advantage of answering this research question with a homogenous, context-relevant population is if individual differences are present, this all but guarantees its presence with more heterogeneous populations. Knowing and quantifying this effect is essential to designing visual displays able to account for workload transitions.

Overall, the motivation to answer these aforementioned research questions is to provide empirically informed design guidelines for environments prone to workload transitions. This study is with a context-relevant population (student Naval aviators) who experienced three different transition rates in the UAV command and control testbed.

Method

Participants

Sixty student Naval aviators participated in this study (50 males, age: *M*=24.5, *SD*=2.3). Participants provided verbal informed consent and the study was approved by the Naval Research Laboratory's Institutional Review Board. Each participant completed three trials (i.e., the testbed scenarios). To assure asymptote performance was being analyzed (as this will be critical to growth curve modeling), participants who were below the 25th percentile during training (i.e., 64% average accuracy across all tasks) only had their 2nd and 3rd trial included in the analysis. There was also data quality threshold for eye tracking data (see details in Chapter 5). The final subset of data included 95 trials from 40 participants, where each transition rate was represented relatively equally, with 32, 34, and 29 trials, of the slow, medium, and fast transitions scenario, respectively (see **Testbed scenarios** section in this chapter for more details).

Experimental setup

The same experimental setup in Chapter 2 was used in Chapter 4. However, the testbed was now presented on a ViewSonic 24" monitor (2560×1440 resolution, 60 Hz refresh rate). Otherwise, all other details from Chapter 2 apply to Chapter 4.

UAV command and control testbed and tasks

The same UAV command and control testbed from Chapter 2 was used for this chapter's research goals.

Testbed scenarios

Workload was manipulated in the same way as Chapter 2. An additional transition rate was created meaning three workload transition rates were tested via three testbed scenarios, i.e., 15-minute missions in the UAV command and control testbed, and were as follows:

1. **Slow transitions scenario**. The number of active UAVs increased steadily from low to high workload. The scenario started at low workload for 100 seconds, and one active UAV was added every 10 seconds until high workload was reached (13-16 active UAVs). The scenario would remain at high workload for 100 seconds, before immediately returning to low workload. This cycle repeated three times for this scenario, meaning there was a total of three periods of low workload and three periods of high workload. The dotted black line in Figure 4.1 depicts the number of simultaneously active UAVs over the course of this testbed scenario.
2. **Medium transitions scenario**. This testbed scenario was the same as Chapter 2. The dashed dark gray line in Figure 4.1 depicts the theoretical number of simultaneously active UAVs over the course of this testbed scenario.
3. **Fast transitions scenario**. This testbed scenario was the same as Chapter 2. The solid

light gray line in Figure 4.1 depicts the number of simultaneously active UAVs over the course of this testbed scenario.

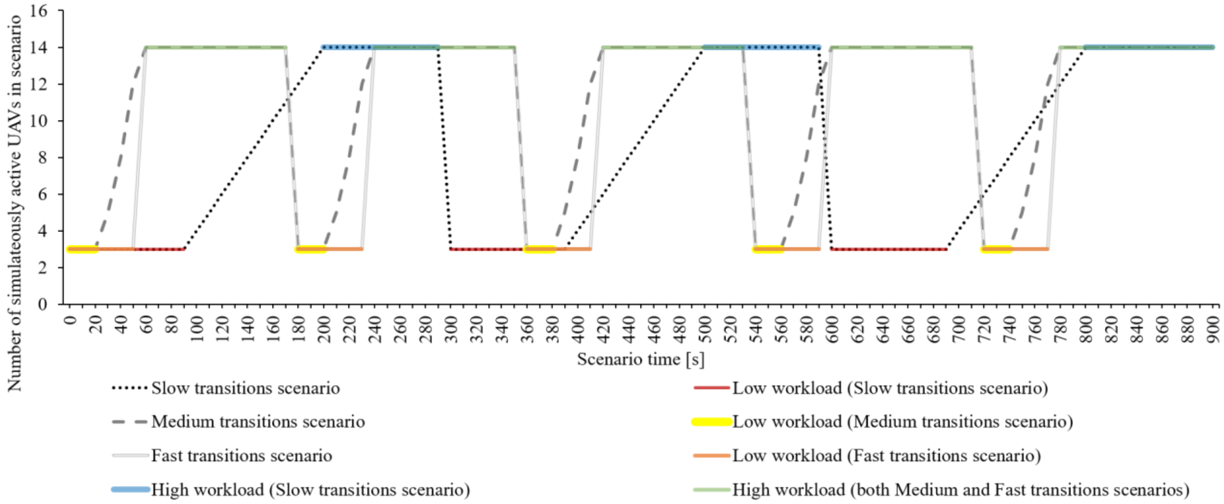


Figure 4.1 The hypothetical number of active UAVs throughout the three testbed scenarios, i.e., transition rates. Each workload period for each testbed scenario is highlighted. Of note, there is only three low and three high workload periods during slow transitions in order to keep testbed scenario length constant

Procedures

This research complied with the APA Code of Ethics and was approved by the Institutional Review Board at the U.S. Naval Research Laboratory. Informed consent was obtained from each participant. Participants then completed a self-paced informational training session and then completed a five-minute training session where 6-10 UAVs were always active. Participants then completed testbed scenarios in a counterbalanced order.

Results

Both the aggregate and growth curve modeling analysis explored how performance trends over time depended on transition rate, workload, and the individual. For both analyses, the dependent variable was always primary task performance, i.e., the target detection task, as this task happened continuously over time. Response time was the time between target onset and its correct detection and accuracy was the percent of correct detections.

Aggregate analysis results

The aggregate analysis consisted of an analysis of variance (ANOVA) on workload period performance for each transition rate. Given the data structure (e.g., different number of trials per participant and different number of workload periods within each transition scenario), restricted maximum likelihood estimation with Satterthwaite's method for degrees of freedom was necessary to accurately conduct the omnibus test and significance was set at α=.05 (Luke, 2017). The ANOVA had one factor with 34 levels, where each level was a combination of transition rate and workload period (i.e., 9 workload periods in slow transitions, 15 workload periods in medium transitions, and 10 workload periods in fast transitions). To determine if there were significant differences between workload periods' response time and accuracy, post hoc tests employed custom linear contrast with Sidak's adjustment for multiple comparisons (Kim, 2015). Adjusted partial eta-squared (η_p^2) is reported as a measure of effect size for the omnibus test (interpretation: very small <0.02, small 0.02-0.13, medium 0.13-0.26, large if >0.26; Cohen, 1988). For pairwise comparisons, Cohen's d for repeated measures (d_{rm}) is reported (interpretation: small=<0.2, medium 0.2-0.5, large >0.8; Lakens, 2013). For all figures, error bars represent the standard deviation of the mean, braces indicate groupings of workload periods, and brackets with an asterisk indicate a significant comparison between two workload periods.

Primary task response time

When analyzing response time, the main effect of the transition rate and workload period was significant (F(33, 1016.6)=32.404, p<.001, adjusted η_p^2=0.5). During slow transitions, response time was slower in the 2nd low workload period (M=3.31 s, SD=0.57 s) than the 1st (M=2.69 s, SD=0.32 s, p=.009, d_{rm}=0.931) and 3rd (M=2.62 s, SD=0.58 s, p=.002, d_{rm}=0.850), suggesting an immediate performance decrement that recovered over time. For medium transitions, response time for the 1st low workload period (M=3.65 s, SD=0.87 s) was slower than the 2nd (M=2.52 s, SD=1.30 s, p<.0001, d_{rm}=0.707), 3rd (M=2.35 s, SD=1.43 s, p<.0001, d_{rm}=0.770), and 4th (M=2.13 s, SD=0.96 s, p<.0001, d_{rm}=1.155). However, the 5th low workload period (M=5.08 s, SD=0.79 s) had the slowest response time compared to all other periods (all p<.0001, d_{rm}=1.155-2.340), suggesting response times were faster with the first instances of medium transitions, but performance eventually returned to initial speeds over time. For fast transitions, the 3rd (M=2.83 s, SD=0.80 s), 4th (M=3.29 s, SD=0.75 s), and 5th low workload periods (M=3.46 s, SD=0.63 s) had significantly slower response times than the 1st (M=1.82 s, SD=0.46 s, all p<.01, d_{rm}=1.073, 1.654, and 2.119 respectively) and 2nd low workload periods (M=2.16 s, SD=1.08 s, all p<.01, d_{rm}=0.491, 0.842, and 1.032 respectively). However, the 5th low workload period was also significantly slower than the 3rd low workload period (p=.02, d_{rm}=0.609), suggesting response time slowed from the onset of the first fast transition. Figure 4.2 details the differences in primary task response time across each low workload period within each transition rate.

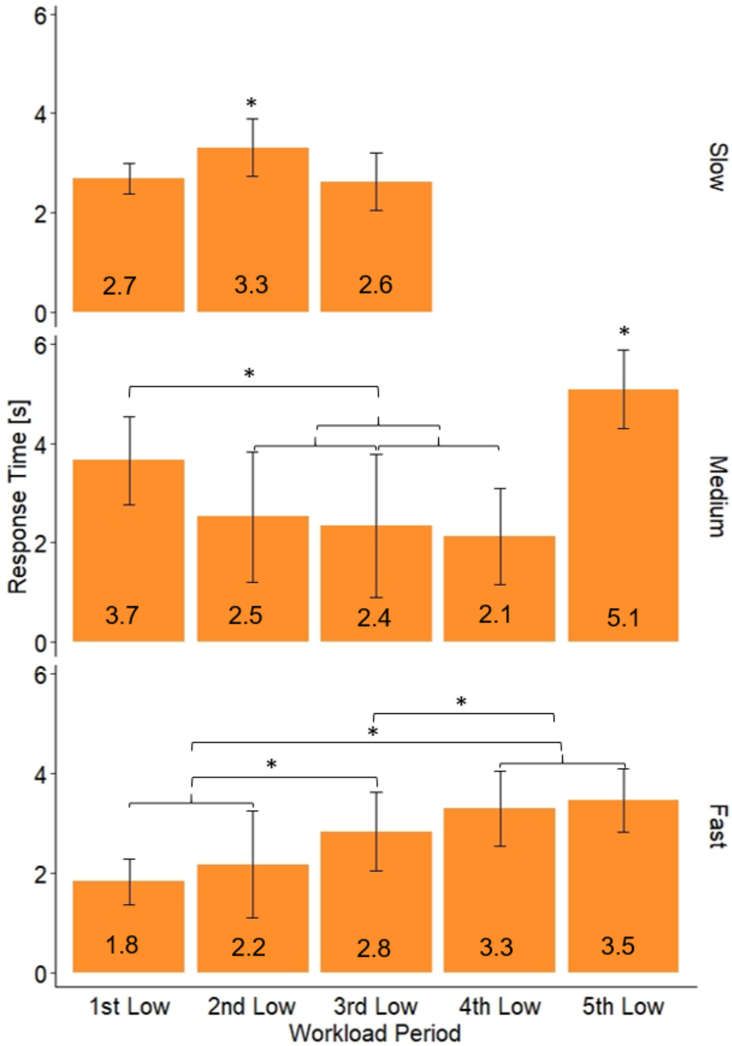


Figure 4.2 Mean primary task response time across the low workload periods in each transition rate. Asterisks (*) denote significant differences between workload periods

For high workload periods, post hoc tests found the 1st high workload period in slow transitions (M=3.22 s, SD=0.40 s) had significantly faster response time than the 2nd (M=3.77 s, SD=0.37 s, p=.046, d_{rm}=1.102). By the 3rd high workload period, response time was closer to initial speeds, but it was not significantly slower or faster than either previous period (M=3.40 s, SD=0.36 s, both p>.05). For medium and fast transitions, there were no significant differences for the response times of high workload periods. Figure 4.3 details the differences in primary task response time across each high workload period within each transition rate.

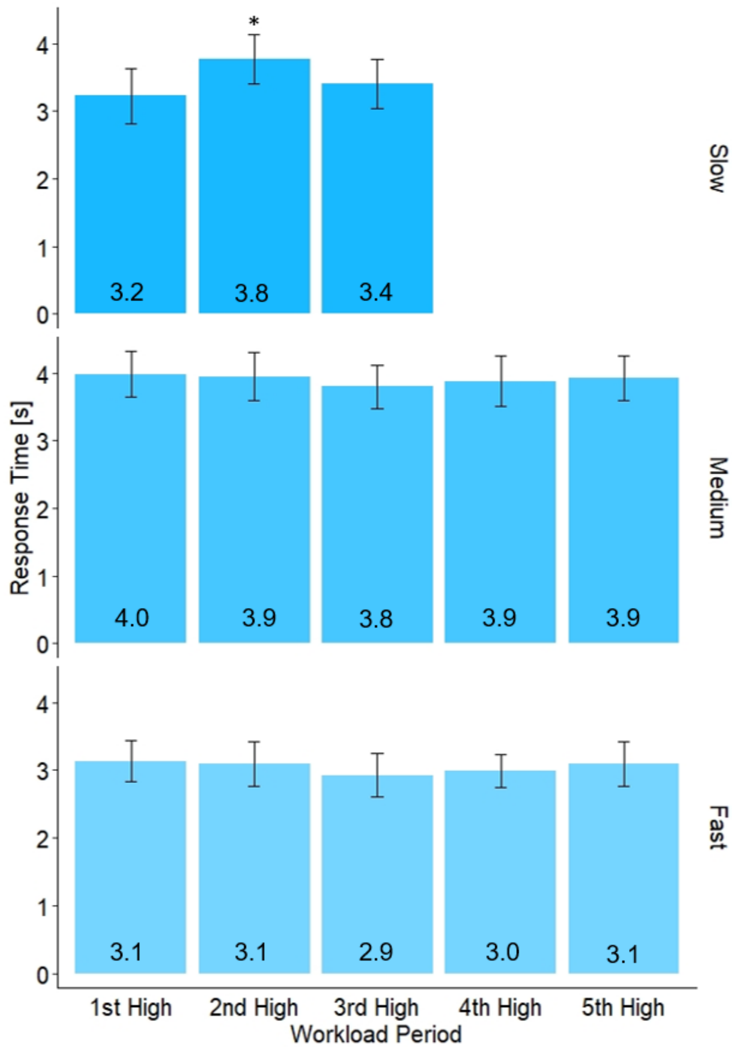


Figure 4.3 Mean primary task response time across the high workload periods in each transition rate. Asterisks (*) denote significant differences between conditions

Primary task accuracy

When analyzing accuracy, the main effect of transition rate and workload period was significant ($F(33,1016.6)=19.422$, $p<.0001$, adjusted $\eta_p^2=0.37$). Post hoc comparisons showed no differences in accuracy across any of the low workload periods for slow transitions (all $p>0.05$).

This was not the case for the medium transitions, as accuracy was higher in the 1st (*M*=92.4%, *SD*=19.9%) and 2nd low workload period (*M*=95.6%, *SD*=18.9%) compared to the 3rd (*M*=68.6%, *SD*=30.6%, both $p<.0001$, d_{rm}=0.632 and 0.734 respectively) and 4th low workload period (*M*=73.5%, *SD*=30.7%, both $p<.0001$, d_{rm}=0.506 and 0.597 respectively). Accuracy recovered by the 5th low workload period (*M*=85.3%, *SD*=24.9%), and was significantly higher than the 3rd low workload period ($p<.0001$, d_{rm}=0.419). For the fast transitions, the 1st low workload period (*M*=91.0%, *SD*=26.5%) had significantly higher accuracy than the 2nd (*M*=73.3%, *SD*=27.0%, $p<0.0001$, d_{rm}=0.469), 3rd (*M*=51.4%, *SD*=29.5%, $p<0.0001$, d_{rm}=0.997), and 4th (*M*=74.5%, *SD*=25.6%, $p=0.002$, d_{rm}=0.449) low workload periods. Again, accuracy significantly drops, but recovers as the 2nd ($p<.0001$, d_{rm}=0.547), 4th ($p<.0001$, d_{rm}=0.588) and 5th low workload periods (*M*=81.6%, *SD*=28.1%, $p<.0001$, d_{rm}=0.741) had significantly higher accuracy than the 3rd low workload period. Figure 4.4 details the differences in primary task accuracy across each low workload period within each transition rate.

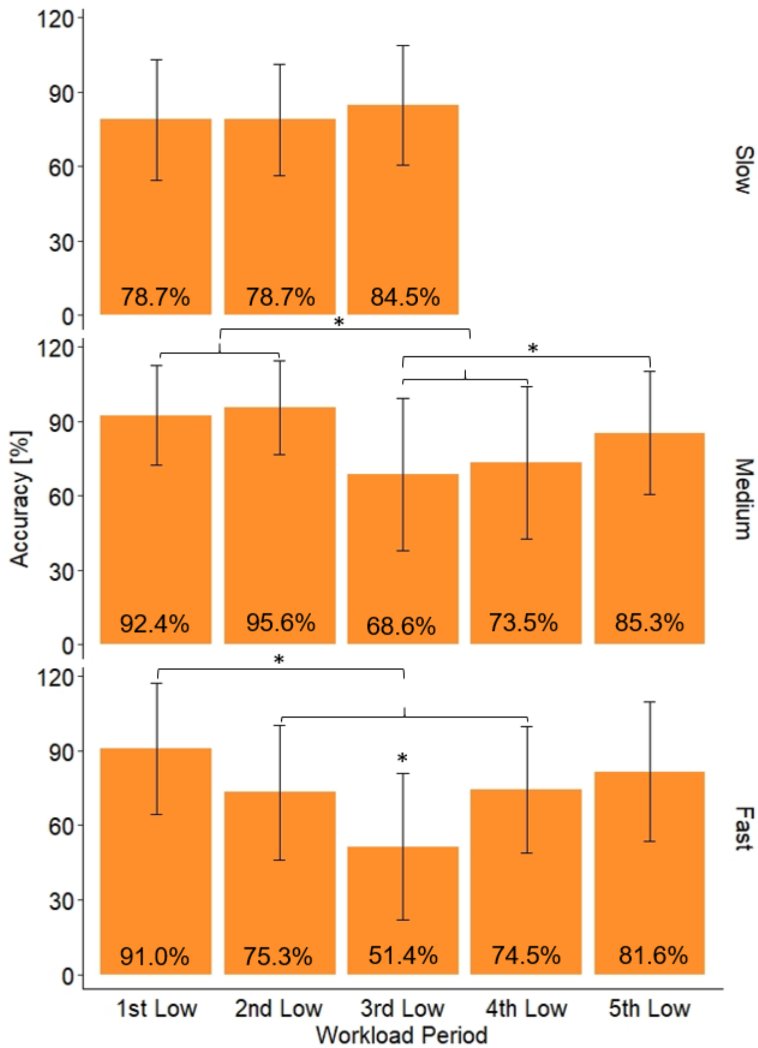


Figure 4.4 Mean primary task accuracy across the low workload periods in each transition rate. Asterisks (*) denote significant differences between conditions

For the accuracy across the high workload periods, there were no significant differences during slow transitions. However, high workload accuracy improved during medium transitions, as the 4th (*M*=69.3%, *SD*=16.0%, *p*=.001, d_{rm}=0.673) and 5th high workload periods (*M*=66.7%, *SD*=17.6%, *p*=.022, d_{rm}=0.539) had significantly higher accuracy than the 1st high workload period (*M*=53.5%, *SD*=17.0%). There was also no significant difference in accuracy across the high workload periods of fast transitions. Figure 4.5 details the differences in primary task accuracy across the high workload periods in each transition rate.

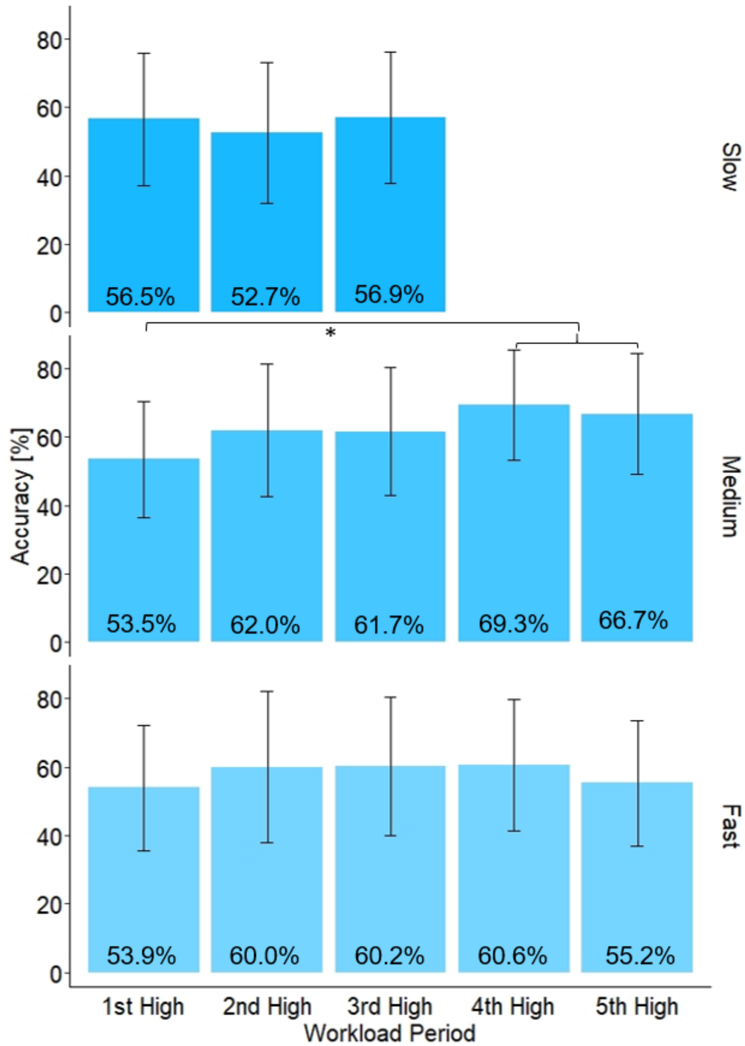


Figure 4.5 Mean primary task accuracy across the high workload periods in each transition rate. Asterisks (*) denote significant differences between conditions

Growth curve modeling

The second analysis consisted of growth curve modeling, where the goal was to determine if and how: (a) performance changed over time (RQ 4.1) and (b) varied across individuals (RQ 4.3). This meant building several growth curve models for both response time and accuracy for each testbed scenario, i.e., slow, medium, and fast transitions. The growth curve models consisted of two types of effects. *Fixed effects* do not vary across the grouping variable of interest, which in this work was always the participant so, the estimated value and significance of a fixed effect applies to all participants. Specifically, the fixed effects estimated the performance trends over time (RQ 4.1). On the other hand, *random effects* capture the variance associated with each parameter, meaning the value for that parameter depends on the participant. In this work, the model's random effects estimate the variance associated with the estimated performance trends over time, which allows for assessing the impact of individual differences (RQ 4.3). Equation 4.1 shows a model estimating values over time t where the intercept (β_{0i}) is a random effect, i.e., a random intercept, because its value depends on person i. However, the predictor (β_1) is a fixed effect because its value does not have this dependency.

$$y_{ti} = \beta_{0i} + \beta_1 Time_{ti} + e_{ti} \tag{4.1}$$

$$where\ \beta_{0i} = \gamma_{00} + U_{0i}$$

The first step in growth curve modeling is to determine the appropriate null model, i.e., the model to build upon when determining the best fit. This consisted of testing if an empty means, random intercept model (Equation 4.2) significantly improved model fit when compared to an empty means only model (Equation 4.3) as determined by a nested likelihood ratio test (LRT) with significance set at α=0.05. Maximum likelihood estimation (ML) was used to fit both models.

$$y_{ti} = \beta_{0i} + e_{ti} \tag{4.2}$$

$$where,\ \beta_{0i} = \gamma_{00} + U_{0i}$$

$$y_{ti} = \beta_0 + e_{ti} \tag{4.3}$$

If the empty means, random intercept model (Equation 4.2) was a significantly better fit, then the intercept of the growth curve model, which in this work represents baseline performance, had a significant variance ($\tau_{U_0}^2$). In practical terms, this means the intercept value depends on the individual participant, so each participant needs his/her own growth curve. Otherwise, the empty means only model (Equation 4.3) served as the null model.

Then, fixed effects of scenario time, i.e., fixed time slopes, were added to the null model. Scenario time was in 10 second increments because a new target appeared every 10 seconds, on average. Fixed linear, quadratic, and cubic time slopes were of interest; cubic was the highest ordered polynomial considered because previous workload transition research observes performance to change trends over time no more than twice over the course of the experimental session (Gluckman et al.,1993; McKendrick & Harwood, 2019), which can be captured by a cubic time slope. Also, higher order polynomials seldom fit human behavioral data (Hoffman, 2015). Each fixed time slope was sequentially added to the model to see if it significantly improved model fit, as assessed via nested likelihood ration tests (LRTs) where significance was set at α=0.05 (Fitzmaurice, Laird, & Ware, 2011; Hedeker & Gibbons, 2006). When a new, higher ordered polynomial fixed time slope was added, all lower ordered polynomials, regardless of significance, were kept to accurately compare model fit and preserve proper interpretation. Equation 4.4 is an example of a growth curve model with a fixed cubic time slope and random intercept, which would be referred to as a fixed cubic, random intercept growth curve model.

$$y_{ti} = \beta_{0i} + \beta_1(Time_{ti}) + \beta_2(Time_{ti})^2 + \beta_3(Time_{ti})^3 + e_{ti} \tag{4.4}$$
$$where, \beta_{0i} = \gamma_{00} + U_{0i}, \beta_1 = \gamma_{10}, \beta_2 = \gamma_{20}, \ \beta_3 = \gamma_{30}$$

Once the best fitting fixed time slope model was determined, every time slope in the model was then sequentially tested as a random effect, i.e., a random time slope. Again, nested LRTs were used to compare model fit, but significance was now set at α=0.10 to protect against Type II errors (Fitzmaurice et al., 2011; Hedeker & Gibbons, 2006). However, if the model became singular when adding random time slopes, i.e., the variance of that time slope was estimated as non-positive, model fit was not assessed as the likelihoods were no longer comparable (Hoffman, 2015, p. 198). Equation 4.5 is an example of a growth curve model with fixed quadratic and cubic time slopes, random intercept, and random linear time slope, which would be referred to as a fixed cubic, random intercept, and random linear growth curve model.

$$y_{ti} = \beta_{0i} + \beta_{1i}(Time_{ti}) + \beta_2(Time_{ti})^2 + \beta_3(Time_{ti})^3 + e_{ti} \tag{4.5}$$
$$where, \ \beta_{0i} = \gamma_{00} + U_{0i}, \ \beta_{1i} = \gamma_{10} + U_{1i}, \ \beta_2 = \gamma_{20}, \ \beta_3 = \gamma_{30}$$

Again, maximum likelihood estimation (ML) method was used to fit each model. Time slopes were scaled to assure model convergence, but rescaled coefficients are presented when practical interpretation is needed, i.e., when presenting the best fitting growth curve model (Bates et al., 2015). RStudio (version 1.3.1093) was used for all analyses: response time was modeled as a normal distribution and built with the *nlme* (Pinheiro et al., 2021) and *lme4* package (Bates et al., 2015; RStudio Team, 2020). Accuracy was modeled as a binomial distribution and built with *GLMMadaptive* package in RStudio to robustly fit models and estimate the standard errors of the fixed effects (Rizopoulos, 2021). The scaled coefficients and fit statistics of each model are reported in Tables 4.1-4.6. In the interest of brevity, not every model that was built is reported. Rather, only the details of the null model, each fixed time slope model, and the random linear time slope model are presented because making the quadratic or cubic time slopes random effects never improved model fit. The final growth curve model, i.e., the best fitting model overall, for each performance metric of each transition rate is interpreted and discussed in the text and highlighted in its respective table.

Response Time: Slow transitions scenario

The empty means, random intercept model was a significantly better model fit than an empty means only model ($\chi^2(1)=18.581$, $p<.0001$). This indicated baseline response time speeds varied significantly across participants ($\tau^2_{U_0}= 0.038$) and an empty mean, random intercept model should be the null model. To determine how response time trended over the course of slow transitions, linear ($\chi^2(1)=0.252$, $p=0.616$), quadratic ($\chi^2(1)=19.901$, $p<.001$), and cubic ($\chi^2(1)= 83.385$, $p<.001$) time slopes were incrementally added to the model and compared. The fixed quadratic and cubic time slopes significantly improved model fit (details of these models in Table 4.1). Including a random linear time slope did not lead to a significantly better model fit (Model 4, $\chi^2(2)=1.742$, $p=0.419$) and all other random time slopes lead to singular model fits, so they were not compared. Therefore, a fixed positive cubic time slope and random intercept model was the best fitting growth curve model of response time during slow transitions (Model 3 in Table 4.1).

Starting response time, i.e., the intercept, was estimated to be 2.4 s; however, the intercept was a significant random effect, meaning the intercept varied significantly across participants (p<.001, $\tau^2_{U_0}$=0.04, estimated range: [2.1-2.7 s]). Response time then changed cubically over time, as it was estimated to increase by 1.2 s during the first 300 seconds, then decrease by 0.5 s during the next 400 seconds, and then increase again by 0.7 s for the remainder of the scenario. Table 4.1 presents each model and its comparative fit. Figure 4.6 shows how the growth curve model estimates response time trends for each participant, specifically showing response time depended on the participant, but response time slowed for all participants over the course of the slow transitions.

Table 4.1 Slow transitions' scaled estimates of each growth curve model fitted for primary task response time. The best fitting model is bolded and highlighted in gray

	Null model (Empty means, random intercept)	Model 1 (Fixed linear, time slope, random intercept)	Model 2 (Fixed linear and quadratic time slopes, random intercept)	**Model 3 (Fixed linear, quadratic, and cubic time slopes, random intercept)**	Model 4 (Fixed linear, quadratic, and cubic time slopes, random intercept and linear time slope)
Intercept (γ_{00})	3.247*** (0.04449)	3.247*** (0.04450)	3.247*** (0.04447)	3.247*** (0.04424)	3.248*** (0.04427)
Fixed linear time slope (γ_{10})	-	0.014 (0.028)	0.505*** (0.113)	2.482*** (0.271)	2.482*** (0.271)
Fixed quadratic time slope (γ_{20})	-	-	-0.506*** (0.113)	-5.695*** (0.658)	-5.690*** (0.657)
Fixed cubic time slope (γ_{30})	-	-	-	3.315*** (0.414)	3.311*** (0.414)
Intercept variance ($\tau^2_{U_0}$)	0.038	0.038	0.038	0.038	0.039
Linear slope variance ($\tau^2_{U_1}$)	-	-	-	-	0.003
Fit Statistics					
LL	-3987.6	-3987.5	-3977.5	-3945.9	-3945.0
χ^2df	-	1	1	1	2
χ^2	-	0.252	19.901	83.385	1.742
p	-	0.616	<0.001***	<0.001***	0.419

Significance codes: '***' indicates p<0.001, '**' indicates p<0.01, '*' indicates p<0.05 '*', '†' indicates p<0.1

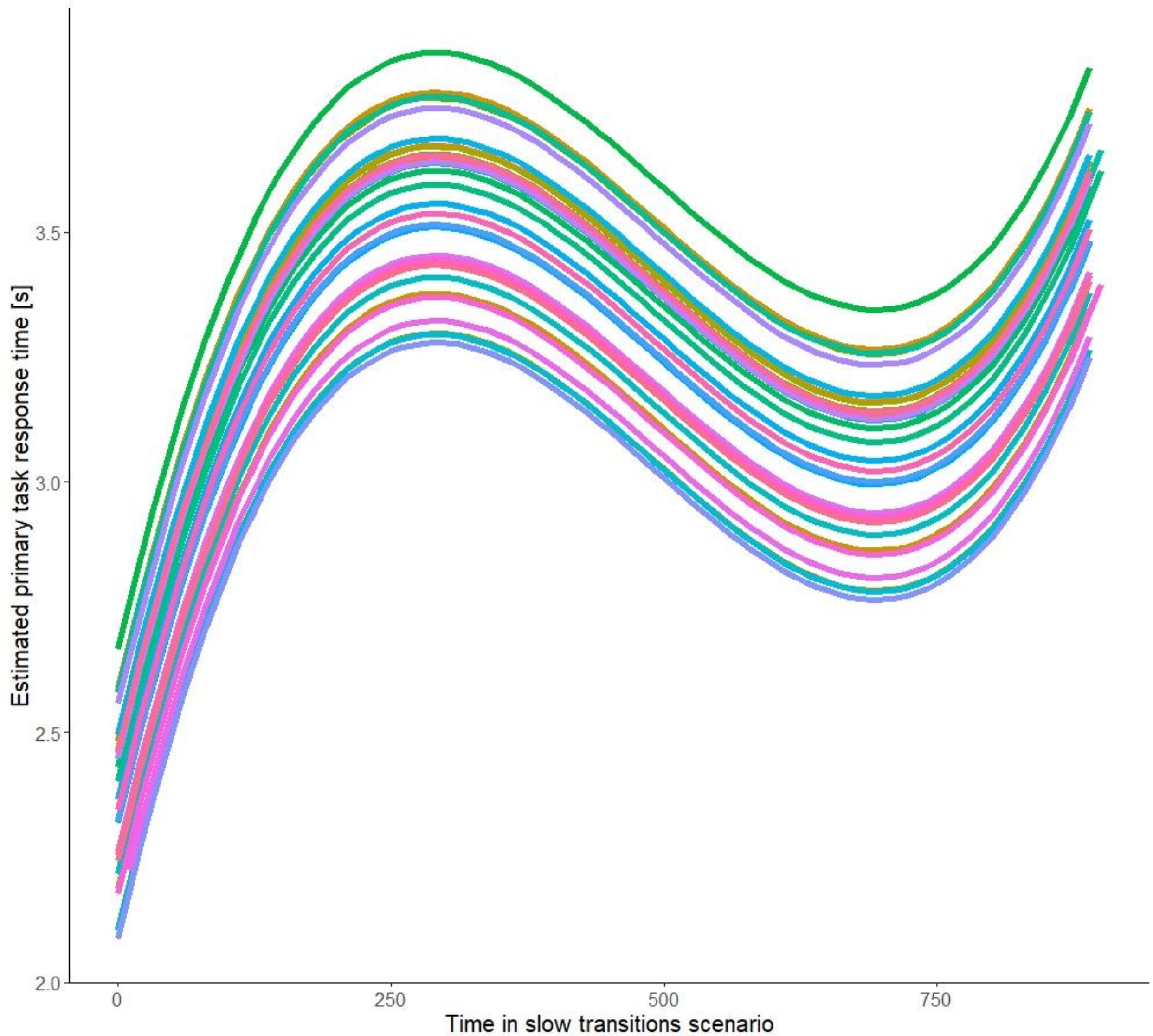


Figure 4.6 Slow transitions' estimated growth curve of primary task response time for each participant. Response time varied across participants, but it was estimated to follow a cubic trend over time

Response time: Medium transitions scenario

The empty means, random intercept model was a significantly better model fit than an empty means only model ($\chi^2(1)$= 22.023, *p*<.0001), meaning participant's baseline response time speeds varied significantly for medium transitions ($\tau^2_{U_0}$= 0.039). Therefore, an empty mean, random intercept model was the null model when adding in fixed time slopes. To determine how response time trended over the course of the medium transitions scenario, linear ($\chi^2(1)= 0.760$, $p=0.383$), quadratic ($\chi^2(1)=7.263$, $p=.007$), and cubic ($\chi^2(1)=1.628$, $p=0.202$) time slopes were incrementally added and compared. A quadratic time slope was the only one to improve model fit (details of these models in Table 4.2). Random linear and quadratic time slopes were added, but both lead to singular model fits, so they could not be compared. Therefore, a fixed positive quadratic time slope with a random intercept was the best fitting growth curve model of response time during medium transitions (Model 2 in Table 4.2).

Starting response time, i.e., the intercept, was estimated to be 4.0 s; however, the intercept was a significant random effect, so starting response time varied significantly across participants ($p<.001$, $\tau^2_{U_0}= 0.042$, estimated range: [3.6-4.3 s]). However, all participants were expected to experience an estimated improvement in response time of 0.2 s during the first 430 seconds of medium transitions, but a decrement of 0.3 s followed for the remainder of the scenario, meaning response time essentially returned to initial speeds over time. Table 4.2 presents each model and its comparative fit. Figure 4.7 shows how the growth curve model estimates response time trends for each participant, specifically showing its speed depended on the participant, but all participants trended similarly over the course of medium transitions.

Table 4.2 Medium transitions' scaled estimates for each growth curve model of primary task response times. The best fitting model is bolded and highlighted in gray

	Null model (Empty means, random intercept)	Model 1 (Fixed linear time slope, random intercept)	**Model 2 (Fixed linear and quadratic time slope, random intercept)**	Model 3 (Fixed linear, quadratic, and cubic time slope random intercept)
Intercept (γ_{00})	3.845*** (0.044)	3.850*** (0.044)	3.850*** (0.044)	3.850*** (0.044)
Fixed linear time slope (γ_{10})	-	0.033 (0.027)	-0.257* (0.107)	0.054 (0.266)
Fixed quadratic time slope (γ_{20})	-	-	0.289** (0.107)	-0.517 (0.641)
Fixed cubic time slope (γ_{30})	-	-	-	0.512 (0.402)
Intercept variance ($\tau^2_{U_0}$)	0.041	0.041	0.042	0.042
Fit Statistics				
LL	-4615.5	-4615.1	-4611.5	-4610.7
χ^2 df	-	1	1	1
χ^2	-	0.760	7.263	1.628
p	-	0.383	0.007**	0.202

Significance codes: '***' indicates $p<0.001$, '**' indicates $p<0.01$, '*' indicates $p<0.05$ '*', '†' indicates $p<0.1$

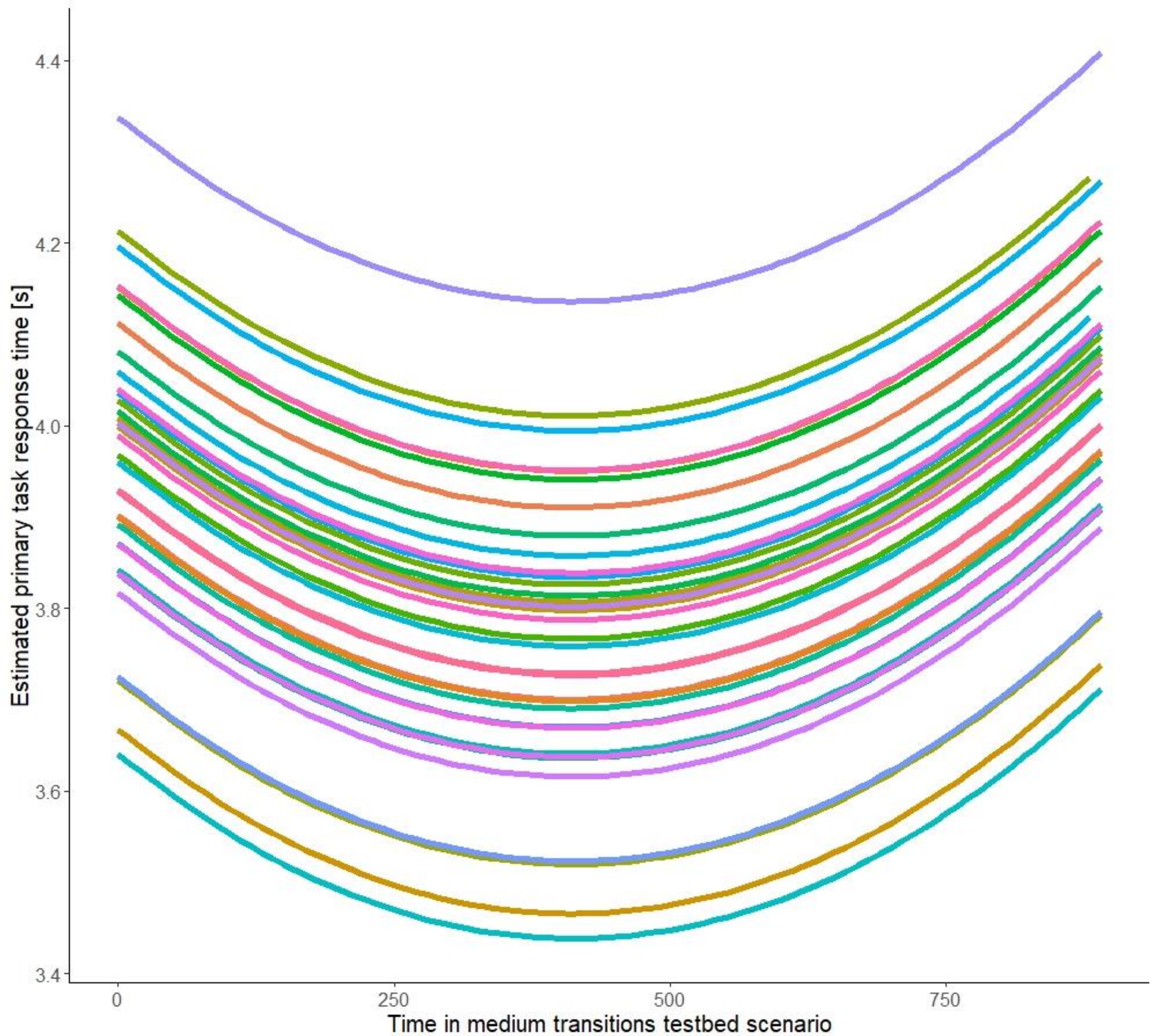


Figure 4.7 Medium transitions' estimated growth curve of primary task response time for each participant. Response time speeds depended on the participant, but it was estimated to follow quadratic trend over time

Response Time: Fast transitions scenario

The empty means, random intercept model was not a significantly better model fit than an empty means only model ($\chi^2(1)$= 2.159, *p*=.142) indicating that response time did not significantly differ amongst participants. Therefore, the null model was an empty means only model. As a follow up, random time slopes were explored, but none significantly improved model fit (all $p>0.05$), confirming individual differences in response time were not present during fast transitions. Therefore, models were fit with the same three fixed polynomial time slopes with no random effects (all details in Table 4.3; linear: $\chi^2(1)= 32.202$, $p<0.0001$; quadratic: $\chi^2(1)= 20.047$, $p<0.0001$; cubic: $\chi^2(1)= 7.160$, $p=.008$). All polynomial time slopes improved model fit, making a fixed cubic time slope with no random effects the best fitting growth curve model of response time during fast transitions (Model 3 in Table 4.3).

Starting response time (i.e., the intercept) was estimated to be 2.27 s, but it would increase by 0.9 s during the first 440 seconds of the scenario. Then, it was estimated to decrease by 0.06 s for the following 280 seconds before increasing again for the remainder of the scenario, ending at an estimated total increase of 0.90 s. Table 4.3 presents each model and its comparative fit and Figure 4.8 shows the best fitting growth curve model is the same for all participants.

Table 4.3 Fast transitions' scaled estimates of each growth curve model of primary task response times. The best fitting model is bolded and highlighted in gray

	Null (Empty means model)	Model 1 (Fixed linear time slope)	Model 2 (Fixed linear and quadratic time slope)	**Model 3 (Fixed linear, quadratic, and cubic time slope)**
Intercept (γ_{00})	2.995*** (0.028)	2.995*** (0.028)	2.995*** (0.028)	2.995*** (0.028)
Fixed linear time slope (γ_{10})	-	0.159*** (0.028)	0.639*** (0.111)	1.309*** (0.273)
Fixed quadratic time slope (γ_{20})	-	-	-0.496*** (0.111)	-2.231*** (0.658)
Fixed cubic time slope (γ_{30})	-	-	-	1.103** (0.412)
Fit Statistics				
LL	-3638.1	-3621.9	-3611.9	-3608.4
χ^2 df	-	1	1	1
χ^2	-	32.202	20.047	7.159
p	-	<0.0001***	<0.0001***	0.008**

Significance codes: '***' indicates p<0.001, '**' indicates p<0.01, '*' indicates p<0.05 '*', '†' indicates p<0.1

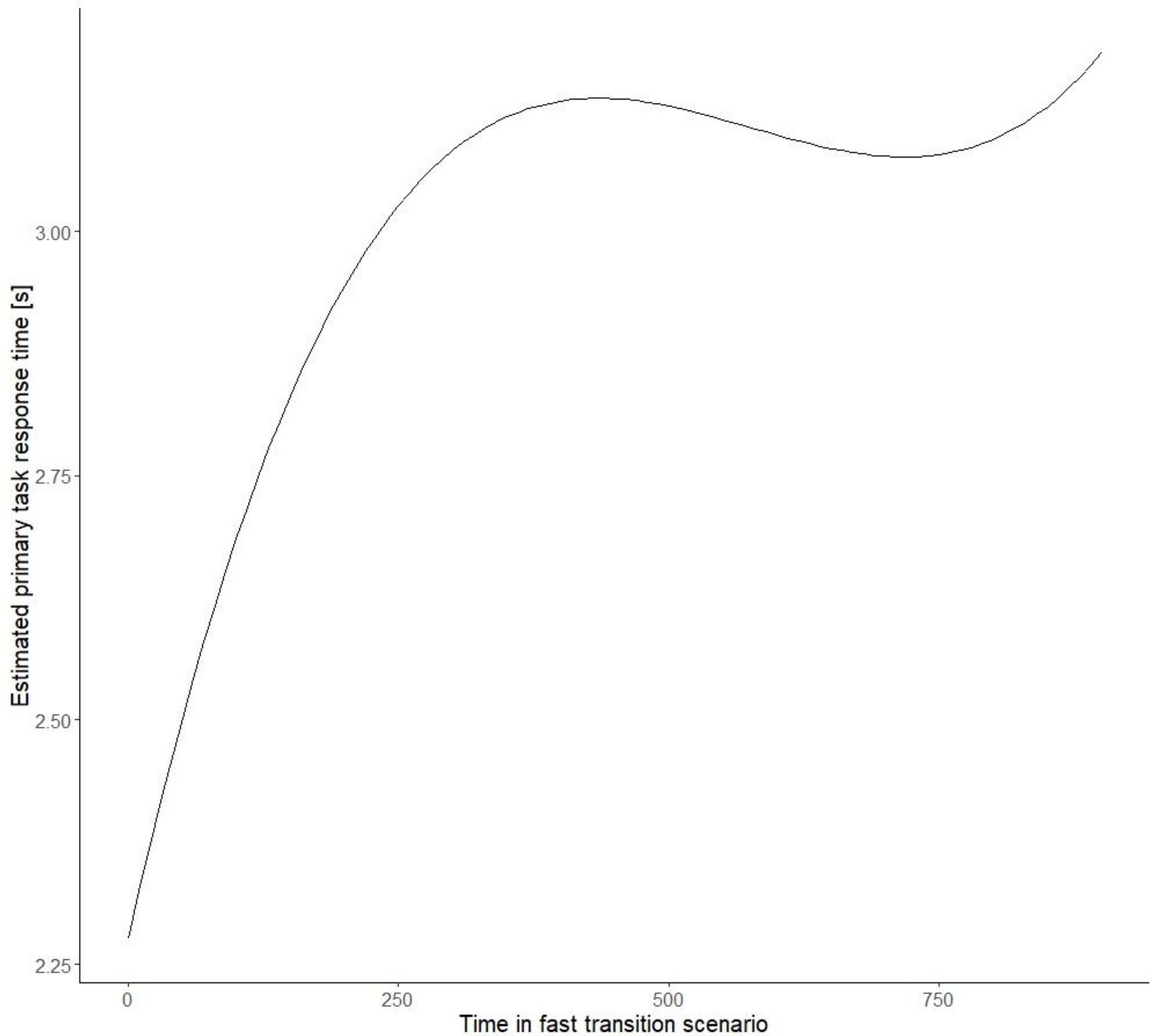


Figure 4.8 Fast transitions' estimated growth curve of primary task response time. A single growth curve applies to all participants because there was no random intercept or time slope

Accuracy: Slow transitions scenario

The empty means, random intercept model significantly improved model fit ($\chi^2(1)$=122.977, p<.0001) meaning accuracy varied significantly across participants ($\tau_{U_0}^2$= 0.20) and an empty means, random intercept model would be the null model. To determine how accuracy trended over time, linear ($\chi^2(1)$= 4.79, p<0.029), quadratic ($\chi^2(1)$= 0.39, p=.530), and cubic ($\chi^2(1)$=66.58, p<0.001) time slopes were incrementally added and compared. Fixed linear and cubic time slopes significantly improved model fit (details of these models in Table 4.4). Adding a random linear, quadratic, or cubic time slope did not improve model fit, i.e., all LRTs produce p>0.10 with the fit statistics from Model 4 in Table 4.4 as an example: ($\chi^2(1)$=0.19, p=.912), meaning how performance trended over time did not significantly depend on the participant. Therefore, a fixed negative cubic time slope and random intercept is the best fitting growth curve model of accuracy during slow transitions (Model 3 in Table 4.4).

Starting accuracy, i.e., the intercept, was estimated to be 87.4%, but it was a random effect so intercept values significantly varied across participants (p<.001, $\tau^2_{U_0}$= 0.21, estimated range: [71.0-92.4%]). However, all participants were estimated to initially decline by 25.6% for the first 300 seconds, somewhat recover by 10.5% for another 360 seconds, before declining again, so that accuracy was 32.7% lower by the end of the scenario. Table 4.4 presents the details of each model and its comparative model fit. Figure 4.9 shows how accuracy rates varied for each participant, yet all trended similarly over the course of slow transitions.

Table 4.4 Slow transitions' scaled estimates of each growth curve model of primary task accuracy. The best fitting model is bolded and highlighted in gray

	Null model (Empty means, random intercept model)	Model 1 (Fixed linear time slope, random intercept)	Model 2 (Fixed linear and quadratic time slope, random intercept)	**Model 3 (Fixed linear, quadratic, and cubic time slope, random intercept)**	Model 4 (Fixed linear, quadratic, and cubic time slope, random intercept and linear time slope)
Intercept (γ_{00})	0.708*** (0.084)	0.718*** (0.084)	0.718*** (0.084)	0.776*** (0.088)	0.776*** (0.088)
Fixed linear time slope (γ_{10})	-	-0.065** (0.023)	-0.147 (0.114)	-2.992*** (0.273)	-3.005 (0.273)
Fixed quadratic time slope (γ_{20})	-	-	0.079 (0.109)	6.829*** (0.608)	6.855 (0.608)
Fixed cubic time slope (γ_{30})	-	-	-	-4.068*** (0.367)	-4.082 (0.367)
Intercept variance ($\tau^2_{U_0}$)	0.201	0.201	0.201	0.201	0.208
Linear slope variance ($\tau^2_{U_1}$)	-	-	-	-	0.0005
Fit Statistics					
LL	-2388.7	-2386.3	-2386.1	-2352.8	-2352.9
χ^2 df	-	1	1	1	2
χ^2	-	4.79	0.39	66.58	0.19
p	-	0.029*	0.530	<0.001***	0.912

Significance codes: '***' indicates p<0.001, '**' indicates p<0.01, '*' indicates p<0.05 '*', '†' indicates p<0.1

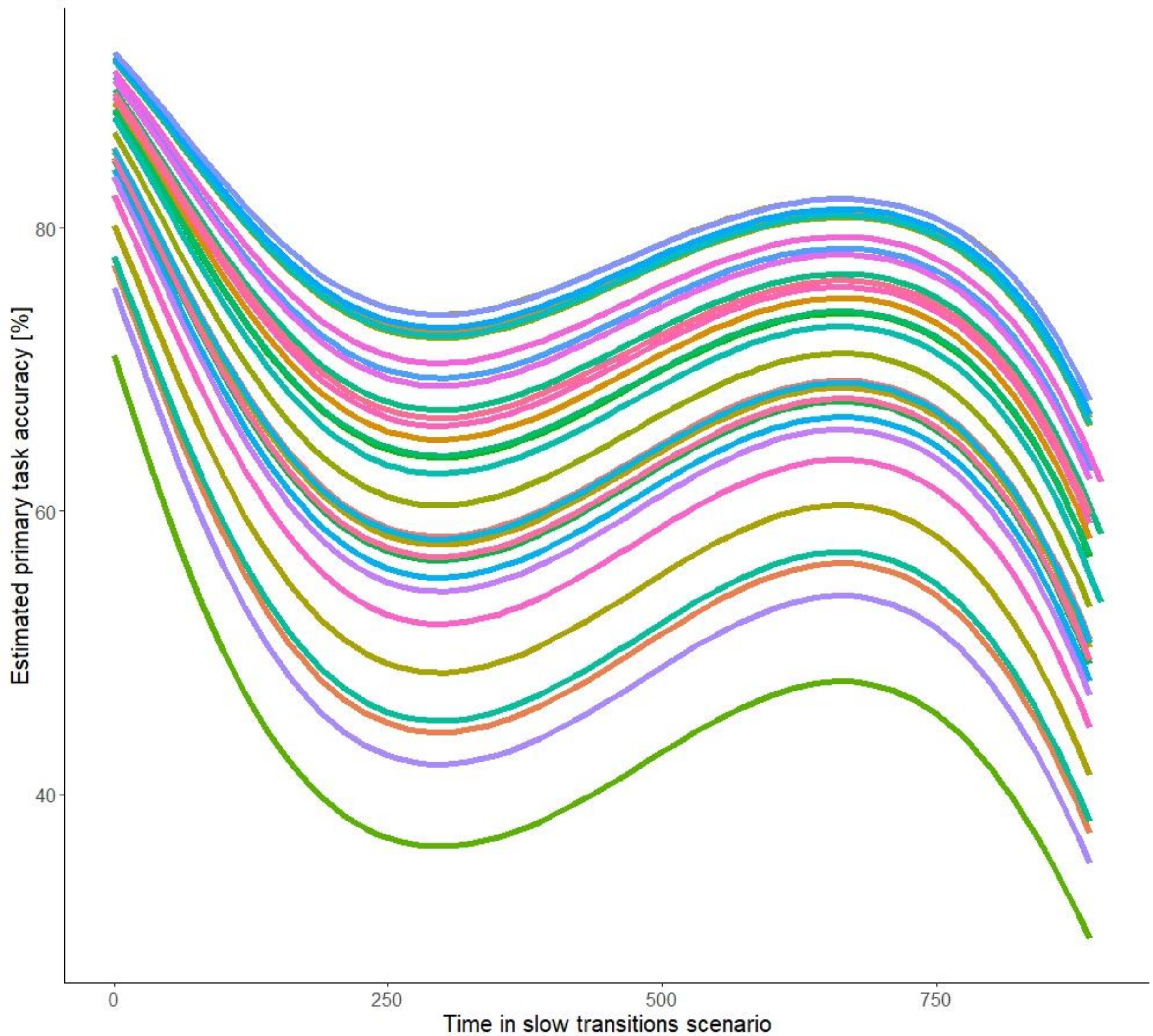


Figure 4.9 Slow transitions' estimated growth curve of primary task accuracy rates for each participant. Accuracy rates depended on the participant, but it was estimated to follow a cubic trend over time

Accuracy: Medium transitions scenario

The empty means, random intercept model was a significantly better fit ($\chi^2(1)$=113.791, *p*<.001) meaning accuracy varied significantly across participants ($\tau^2_{U_0}$=0.14) and an empty means, random intercept model was the null model. To determine how accuracy trended over time, fixed linear ($\chi^2(1)$= 32.3, p<.001), quadratic ($\chi^2(1)$=0, p=.95), and cubic ($\chi^2(1)$= 4.09, p=.043) time slopes were incrementally added to the model. Fixed linear and cubic time slopes improved model fit (details of these models in columns 2 and 4 Table 4.5). Additionally, including a random linear time slope significantly improved model fit ($\chi^2(1)$= 4.74, p=0.09) suggesting performance over time did not trend in the same way for all participants. Therefore, a fixed negative cubic time slope, random intercept and random linear time slope is the best fitting growth curve model of accuracy during medium transitions (Model 4 in Table 4.5).

Starting accuracy (i.e., the intercept) was estimated to be 65.5%, but the intercept was a significant random effect, meaning participants' intercept value varied significantly (p<.001, $\tau^2_{U_0}$= 0.15, estimated range: [52.2-76.3%]). Over time, accuracy was estimated to decrease by 2.5% in the first 170 seconds of the scenario for all participants. However, the significant random linear time slope shows the trends afterwards differed across participants. For reference, 18% of the participants were estimated to decline in accuracy whereas 82% were estimated to improve over the course of medium transitions. Table 4.5 presents the details of each model and its comparative fit. Figure 4.10 shows how accuracy rates and trends depended on each participant for medium transitions.

Table 4.5 Medium transitions' scaled estimates of each growth curve model of primary task accuracy. The best fitting model is bolded and highlighted in gray

	Null model (Empty means, random intercept)	Model 1 (Fixed linear time slope, random intercept)	Model 2 (Fixed linear and quadratic time slope, random intercept)	Model 3 (Fixed linear, quadratic, and cubic time slope, random intercept)	**Model 4 (Fixed linear, quadratic, and cubic time slope, random intercept and linear time slope)**
Intercept (γ_{00})	0.728*** (0.070)	0.726*** (0.070)	0.727*** (0.070)	0.732*** (0.070)	0.734*** (0.070)
Fixed linear time slope (γ_{10})	-	0.145*** (0.030)	0.139 (0.075)	-0.382 (0.245)	-0.387 (0.240)
Fixed quadratic time slope (γ_{20})	-	-	0.006 (0.072)	1.317* (0.603)	1.334* (0.595)
Fixed cubic effect of time (γ_{30})	-	-	-	-0.821* (0.383)	-0.826* (0.381)
Intercept variance ($\tau^2_{U_0}$)	0.143	0.145	0.145	0.145	0.146
Linear slope variance ($\tau^2_{U_1}$)	-	-	-	-	0.009
Fit Statistics					
LL	-3012.5	-2996.4	-2996.4	-2994.3	-2991.9
χ^2 df	-	1	1	1	2
χ^2	-	32.3	0.01	4.09	2.31
p	-	<0.001	0.955	0.043*	0.093†

Significance codes: '***' indicates p<0.001, '**' indicates p<0.01, '*' indicates p<0.05 '*', '†' indicates p<0.1

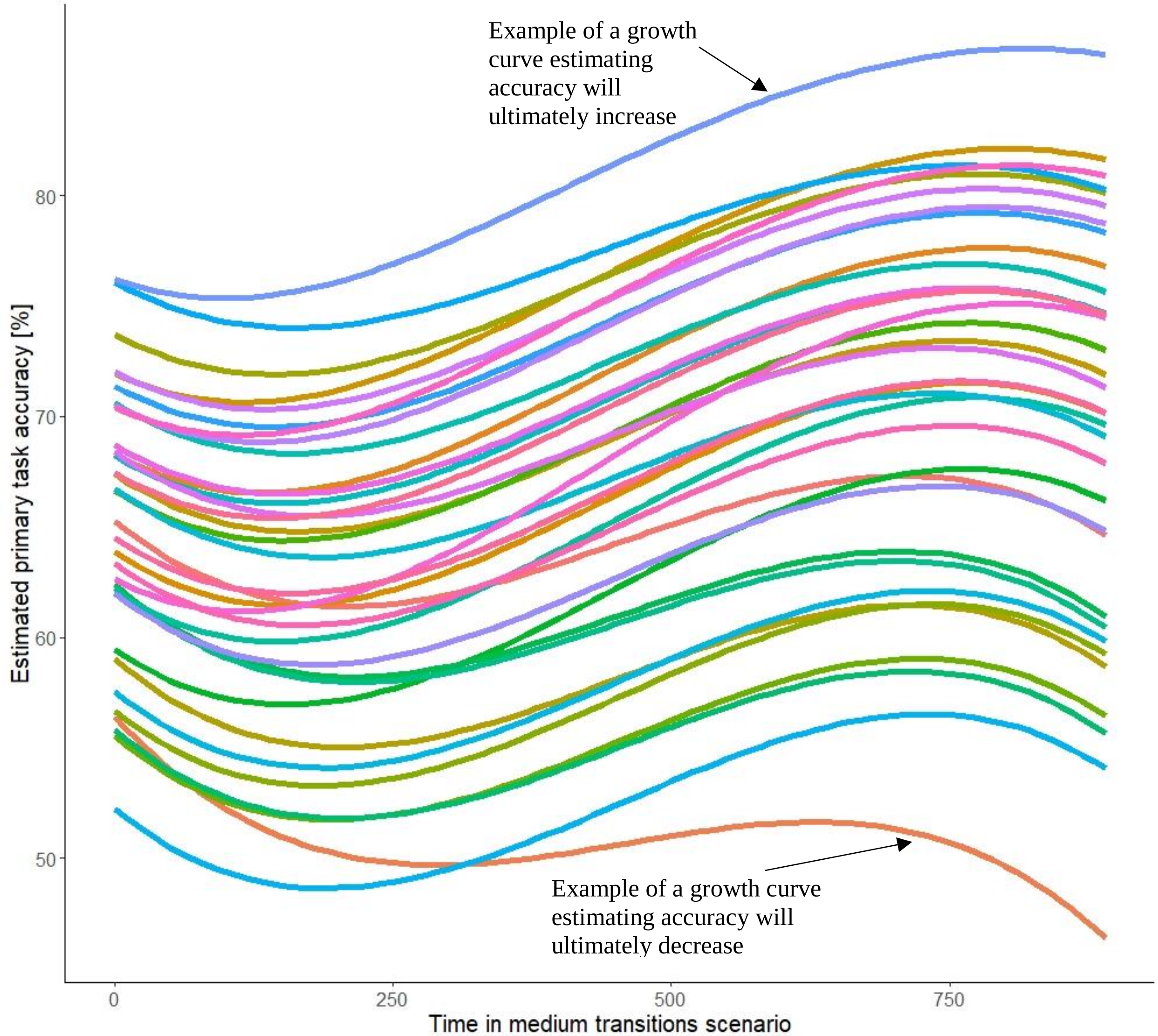


Figure 4.10 Medium transitions' estimated growth curve of primary task accuracy for each participant. As pictured, baseline accuracy rates and accuracy rates over time varied across all participants. All participants followed a cubic trend, but performance could improve, remain the same, or get worse by the end of the medium transitions scenario

Accuracy: Fast transitions scenario

The empty means, random intercept model significantly improved model fit ($\chi^2(1)=116.576$, $p<.0001$) meaning accuracy varied significantly across participants ($\tau_{U_0}^2 = 0.171$)

and an empty means, random intercept model was the null model. To determine how accuracy trended over time, linear ($\chi^2(1)=1.11$, $p=.291$), quadratic ($\chi^2(1)=0.48$, $p=.490$), and cubic ($\chi^2(1)=16.27$, $p<.001$) time slopes were incrementally added to the model. The fixed cubic time slope significantly improved model fit (details of Model 3 Table 4.6). However, no random time slope improved model fit, i.e., all $p>0.10$, review the fit statistics of Model 4 as an example: ($\chi^2(1)=0.19$, $p=.911$). Therefore, a negative cubic time slope with a random intercept was the best fitting growth curve model of accuracy during fast transitions (Model 3 in Table 4.6).

Starting accuracy, i.e. the intercept, was estimated to be 72.5%. However, the intercept was a significant random effect, meaning starting accuracy varied significantly amongst participants ($p<.001$, $\tau^2_{U_0}=0.172$, estimated range: [50.0-85.6%]). As for how accuracy changed over time, the best fitting model predicted accuracy would initially decline by 10.7% for the first 270 seconds, increase by 5.2% for another 360 seconds, before declining again for a total decrease of 15.9%. Table 4.6 presents each model's details and its comparative fit. Figure 4.11 shows how the growth curve model estimates accuracy to depend on each participant, but all participants followed the same trend over the course of fast transitions.

Table 4.6 Fast transitions' scaled estimates of each growth curve model of primary task accuracy. The best fitting model is bolded and highlighted in gray

	Null model (Empty means, random intercept)	Model 1 (Fixed linear time slope, random intercept)	Model 2 (Fixed linear and quadratic time slope, random intercept)	**Model 3 (Fixed linear, quadratic, and cubic time slope, random intercept)**	Model 4 (Fixed linear, quadratic, and cubic time slope, random intercept and linear time slope)
Intercept (γ_{00})	0.583 (0.082)	0.584*** (0.082)	0.584*** (0.082)	0.598*** (0.081)	0.598*** (0.081)
Fixed linear time slope (γ_{10})	-	-0.029 (0.027)	0.049 (0.100)	-1.127*** (0.229)	-1.129*** (0.228)
Fixed quadratic time slope (γ_{20})	-	-	-0.078 (0.095)	2.805*** (0.495)	2.804*** (0.496)
Fixed cubic time slope (γ_{30})	-	-	-	-1.772*** (0.278)	-1.770*** (0.279)
Intercept variance ($\tau^2_{U_0}$)	0.171	0.171	0.171	0.172	0.175
Linear slope variance ($\tau^2_{U_1}$)	-	-	-	-	0.0009
Fit Statistics					
LL	-2431.1	-2430.5	-2430.3	-2422.2	-2422.1
χ^2 df	-	1	1	1	2
χ^2	-	1.11	0.48	16.27	0.19
p	-	0.291	0.489	<0.001***	0.911

Significance codes: '***' indicates p<0.001, '**' indicates p<0.01, '*' indicates p<0.05 '*', '†' indicates p<0.1

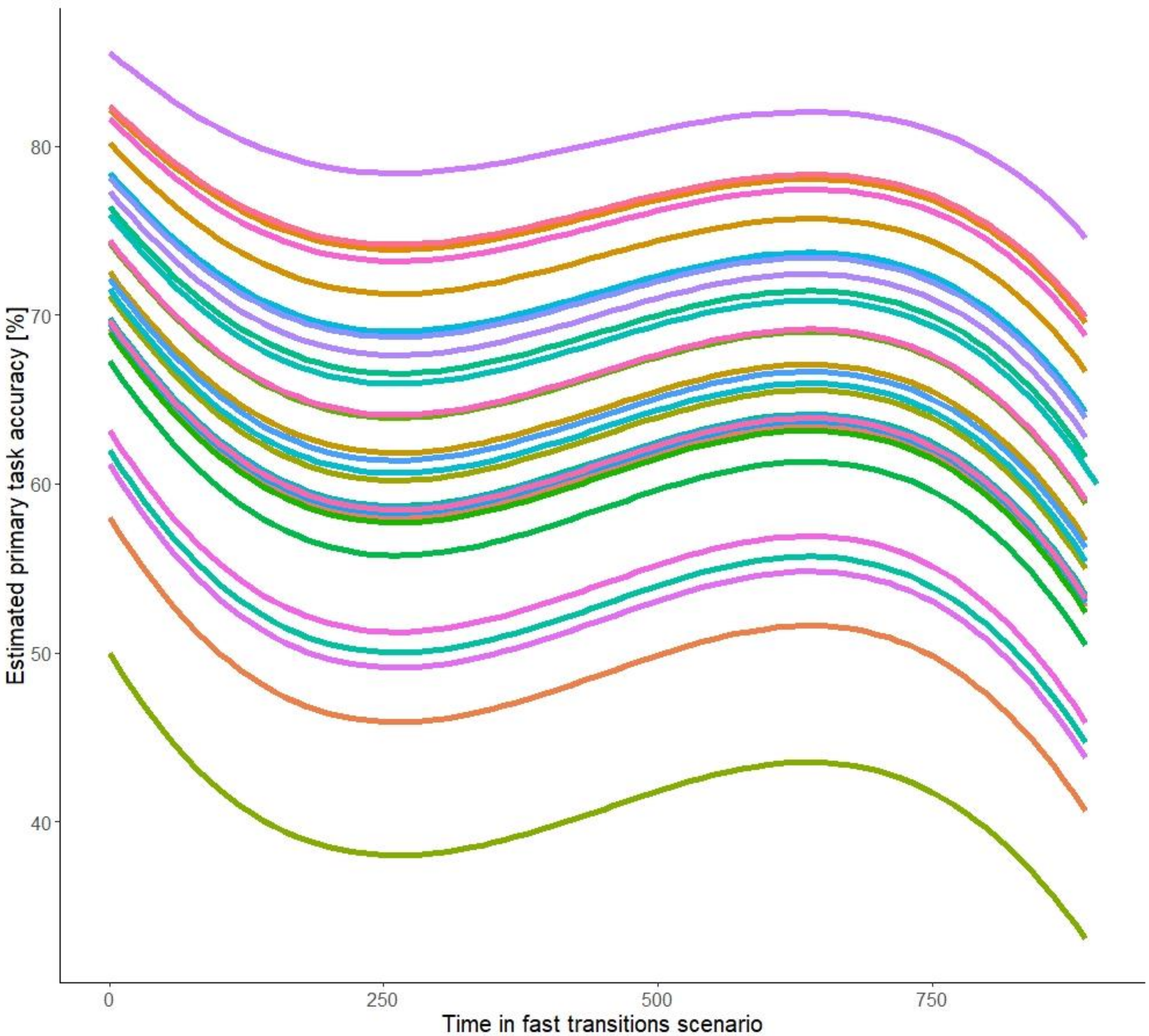


Figure 4.11 Fast transitions' estimated growth curve of primary task accuracy for each participant. Accuracy rates depended on the participant, but it was estimated to follow a negative cubic trend

Table 4.7 summarizes the present results. The table focuses on the general takeaways from each analysis and how they address each research question.

Table 4.7 Summary of all the performance results. Red text highlights where the interpretation of the two analyses agree

Transition rate	**ANALYSIS**	
	Aggregate analysis (Current analysis used by most workload transition studies)	**Growth curve modeling**
Slow	Response time: Initially slows, but eventually recovers to starting speeds for low and high workload periods.	Response time: Baseline speeds differ across individuals, but all individuals follow a *positive cubic trend*, i.e., response time initially slows, slightly improves, and then slows again.
	Accuracy: Not significantly different over the course of the scenario.	Accuracy: Baseline rates differ across individuals, but all individuals follow a *negative cubic trend*, i.e., accuracy initially declines, slightly improves, and then declines again.
Medium	Response time: For low workload periods, it initially speeds up before slowing to its original speed by the end of the scenario. For high workload periods, it did not change over the course of the scenario.	Response time: Baseline speeds differ for each individual. For all participants, response times follow a *positive quadratic trend*, i.e., it initially speeds up but then slows to original speeds by the end of the testbed scenario.
	Accuracy: For low workload periods, rates decline initially, but there is some recovery by the end of the scenario. For high workload periods, accuracy rates improve over the course of the scenario.	Accuracy: Baseline rates and changes in rates over time differ for each individual. Although all participants show a negative cubic trend, **18%** show an eventual **decline** in accuracy by the end of the scenario whereas **82%** show an **improvement**.
Fast	Response time: For low workload periods, it slows continuously over time. For high workload periods, it did not change over the course of the scenario.	Response time: Baseline speeds and changes in speed over time were similar for all individuals as only a single growth curve with a *positive cubic trend*, is needed for all participants. It estimates response times initially slowing, and then showing some evidence of speeding up, before beginning to slow again by the end of the scenario.
	Accuracy: For the low workload periods, rates immediately decline but ultimately improved to previously observed rates. For high workload periods, accuracy rates did not significantly differ over the course of the scenario.	Accuracy: Baseline rates differ for each individual, but all individuals follow a *negative cubic trend*, i.e., accuracy initially declines, slightly improves, yet eventually declines again. The trend was less pronounced, i.e., the slopes were less negative than the cubic trend seen with accuracy during slow transitions.

Discussion

The goal of this work was to understand how transition rate, workload level, and the individual impact performance trends over time for both response time and accuracy when there were multiple instances of workload transitions. Answering each research question builds upon the current knowledgebase and provides design guidelines for environments expected to experience workload transitions.

RQ 4.1: Does transition rate affect performance trends over time?

For RQ 4.1, we expected performance to be different between the transition rates with slow transitions leading to the worst performance trends over time. When reviewing results from both analyses, this was partially supported, as performance was indeed different between the three transition rates. However, the aggregate analysis showed slow transitions had some of the most stable performance trends over time across both low and high workload, even if response time and accuracy was worst compared to the other transition rates. The aggregate analysis also showed that accuracy improves over time for the high workload periods of medium transitions; however, there was a large initial drop in accuracy during low workload periods, while response time remained stagnant throughout. Fast transitions also experienced this large, initial drop in low workload accuracy, but with faster initial response time. Growth curve modeling showed baseline performance was more dependent on the individual, not transition rate. Growth curve modeling also projects performance across participants to worsen over time with slow and fast transitions, but improve or remain stagnant with medium transitions.

Overall, both analyses showed that transition rate impacted performance, but the conclusions from each analysis differed, which is not surprising due to their different approaches. The aggregate analysis measured performance trends, on average, over time within low or high workload periods. Growth curve modeling measured general performance trends over time, independent of workload, and assessed how they differed across individuals. Per these interpretations, performance during workload transitions is dependent on several factors: the workload level, the elapsed time, the performance measure, and the individual. Therefore, it is essential to synthesize the two results to provide goal-oriented and nuanced design guidance to assist operators experiencing workload transitions.

RQ 4.2: Which theoretical explanation—i.e., resource depletion or effort regulation—accounts for the observed over time performance trends?

The results from growth curve modeling, specifically the significant fixed polynomial time slopes, (e.g., linear, quadratic, cubic time slopes), substantially increased the understanding of the applicability of the two main theoretical explanations in the workload transition literature, which addresses RQ 4.2. Under the resource depletion explanation—i.e., workload transitions initially deplete resources, causing performance to suffer, but recovery is possible during low workload—we expected the best fitting growth curve model to have a fixed quadratic time slope, specifically one where performance initially worsens, but then improves. Gluckman et al. (1993) informally observed a quadratic performance trend in their results when developing this theoretical explanation. Under the effort regulation explanation, i.e., workload transition performance is dependent on the individual adjusting and deploying mental resources effectively to the present workload, we expected the best fitting growth curve model to have a random linear time slope as performance depended on the individual appraising and recruiting resources.

However, the best fitting growth curve models were often neither quadratic nor linear, rather they were cubic, i.e., performance was projected to decline initially, improve for some duration of time, before declining again. This suggests both theoretical explanations may have been applicable throughout each scenario, similar to how previous work finds each theoretical explanation depends on the environmental context (Matthews & Desmond, 2002; Ungar et al., 2005). Specifically, resource depletion accounted for the initial performance trends: workload transitions initially deplete resources, so performance initially suffers. Although resource depletion states performance can recover, it stipulates it is only possible during low workload. However, growth curve modeling consistently projected performance to recover during periods of low *and* high workload. This better supports the effort regulation explanation, as participants may have been able to eventually appraise and recruit resources when workload transitioned to maintain improved performance.

The significant random linear time slope estimated for the growth curve model of medium transition accuracy also supported the effort regulation explanation as it contends addressing the dynamic needs in the environment will vary across individuals' ability to appraise the environment and attend to its needs effectively. When performance declines again, as seen for slow and fast transitions, it becomes unclear which explanation better applies. Resources may have been too depleted to maintain the newly improved performance, which would support the resource depletion explanation, *or* the occurrence of multiple transitions made evaluating the environment too challenging to recruit and deploy resources effectively, which aligns with the effort regulation explanation. Further research should address this ambiguity, as multiple workload transitions are expected to be prevalent in a variety of contexts. This research shows the value in using novel methods to build upon the current theory when it struggles to fully explain observed performance trends. In addition, new analysis approaches can also be used to simultaneously account for factors that are not central, but potentially impactful, in a given workload transition experiment. For example, the growth curve modeling results contend baseline performance and performance over time often depended on the individual to some extent, across transition rates, which addressed our third research question.

RQ 4.3: Does the individual affect performance trends over time?

For RQ 4.3, we expected performance to vary significantly across individuals for all transition rates, but particularly during fast transitions. Our hypothesis was mostly supported as all but one best fitting growth curve model showed that baseline performance, i.e., the intercept, varied across participants, yet only one best fitting growth curve model showed performance trending differently over time across participants, i.e., a random linear time slope. Performance being the most variable with medium transitions, actually further supports the effort regulation explanation because this explanation expects performance to vary per the individual (Hockey et al., 1997). Performance variability is expected in these types of environments (Cummings et al., 2019; Muhs et al., 2017) and is sometimes even accounted for a priori (Bowers et al., 2014; McKendrick & Harwood, 2019). However, the observed variability is particularly notable from the homogenous population of student Naval aviators. When previously studying individual differences within workload transitions, research has relied on extreme group design methods, such as comparing subpopulations that are extreme ends of a given measure (Cox-Fuenzalida, et al., 2004, 2006). Recently, this method has been found to be problematic for both experimental and statistical reasons (Tsukahara et al., 2016), but growth curve modeling is not subject to those shortcomings because it is not relying on a priori stratification nor pairwise comparison methods to detect individual differences. Future workload transition research will need to appropriately account for the individual to provide meaningful design guidance, especially for populations more diverse than the one used in this study.

Synthesizing the present results for design guidance

In this chapter, we found:

1. Performance trends differed across transition rates, but the implications of those differences depended on the analysis method.
2. A hybrid of the two dominating theoretical explanations best explained the performance trends. Specifically, workload transition performance was subject to resource depletion *until* effective strategies were discerned by the individual, which the ability to do so depended on the transition rate and the individual.
3. Individuals mostly varied in baseline performance, but individual trajectories also differed depending on the transition rate.

Although both analyses are informative, growth curve modeling addresses substantial research gaps (e.g., the dynamic applicability of theoretical explanations within the same experimental setting, quantifying the impact of individual differences, etc.). Ideally, designing systems that experience workload transitions would continuously take aspects of the environment and individual in to account because the current results suggest that is more likely necessary than not.

However, design guidance from the current results are as presented in Table 4.8 and include foreseeable caveats and tradeoffs.

Table 4.8 Design guidance for workload transitions as a function of performance goal

Performance goal	**Recommended transition rate**	**Justification**	**Potential Tradeoffs**	**Domain Example**
If the goal is consistent and predictable low and high workload performance…	Slow transitions	Performance during slow transitions is more dependent on workload than time, and this will most likely apply to all individuals	Baseline performance is worse in both workload levels than faster transition rates and it does not show the potential to improve over time	Strike missions because straying from expected performance when workload transitions is potentially fatal. These missions can offset the cost in general performance by dispatching additional manpower, delaying engagement, etc. Also, performance-based adaptive automation would be more successful with slow transitions, as the performance differences between low and high workload were discernable and consistent over time (Feigh et al., 2012).
If the goal is for performance to generally improve over time as workload transitions…	Medium transitions	Medium transitions were the only ones that showed the potential to improve over time for both performance measures and across both workload levels.	Potentially may cause an initial, but recoverable performance cost and these benefits might depend on the individual.	Search and rescue missions, as performance needs to improve as time progresses due to the probability of detecting rescues gets harder with time, regardless of present workload. These missions can offset the cost in initial accuracy providing more resources at the beginning of mission and then reducing over time accordingly.
If the goal is faster and predictable performance trends across individuals over time…	Fast transitions	If it's preferable for participants to have similar performance trends, fast transitions may be best as the results indicated fast transitions had the fastest response times and least dependence on the individual.	Potentially there is an initial, but recoverable, cost in accuracy and response times may also slow over time	Sustained reconnaissance missions, as performance needs to be as quick as possible especially in the beginning of the mission. These missions can offset the cost of initial accuracy by weighting the results from earlier parts of the mission differently (O'Rourke, 2006).

Conclusion

Although the impact of this work was threefold, it was not without limitations. Most importantly, future work should investigate how to design for the individual in environments where workload transitions differently as this original investigation has found individual differences to be a significant factor. Although the provided design guidance accounts for this, it is now necessary to discern what specific abilities or strategies individuals rely upon to manage workload transitions to increase the generalizability of design guidance.

Second, although the sample size of participants was relatively large for workload transition research, our sample size may have been underpowered to detect all significant random time slopes, especially higher ordered ones (Astivia et al., 2019). Although a larger sample size was planned (e.g., ~100 participants), the COVID-19 pandemic limited data collection. One random time slope was detected, so the present sample size may have been sufficient, but future work should validate the current results and plan future work accordingly.

Finally, future work needs to include online cognitive measures for both theoretical and practical applications. For example, one psychophysiological metric that has found to be reliable in discerning mental strategies is eye tracking. For example, scan-based metrics (Moacdieh et al., 2020) capture visual attention strategy, which is ideal for understanding the explain effort regulation. Growth curve modeling can readily include these measures, so the next chapter will synthesize the performance and scan-based metrics together to shed light on their direct relationship. Overall, future workload transition research needs to consider including growth curve modeling in their studies, even if only as a follow-up analysis, as it has shown the ability to clarify the cause of ambiguous findings, theoretical explanations, and design guidance.

References

Astivia, O. L. O., Gadermann, A., & Guhn, M. (2019). The relationship between statistical power and predictor distribution in multilevel logistic regression: A simulation-based approach. *BMC Medical Research Methodology*, *19*(1). https://doi.org/10.1186/s12874-019-0742-8

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bowers, M. A. (2013). *The effects of workload transitions in a multitasking environment*. [Master's thesis, University of Dayton]. OhioLINK.

Bowers, M. A., Christensen, J. C., & Eggemeier, F. T. (2014). The effects of workload transitions in a multitasking environment. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 220–224. https://doi.org/10.1177/1541931214581046

Cohen, J. (1988). Analysis of variance. In *Statistical power analysis for the behavioral sciences* (pp. 273–406).

Cox-Fuenzalida, L.-E. (2007). Effect of workload history on task performance. *Human Factors*, *49*(2), 277–291. https://doi.org/10.1518/001872007X312496

Cox-Fuenzalida, L.-E., Angie, A., Holloway, S., & Sohl, L. (2006). Extraversion and task performance: A fresh look through the workload history lens. *Journal of Research in Personality*, *40*(4), 432–439. https://doi.org/10.1016/j.jrp.2005.02.003

Cox-Fuenzalida, L.-E., Beeler, C., & Sohl, L. (2006). Workload history effects: A comparison of sudden increases and decreases on performance. *Current Psychology*, *25*(1), 8–14.

https://doi.org/10.1007/s12144-006-1012-6

Cox-Fuenzalida, L.-E., Swickert, R., & Hittner, J. B. (2004). Effects of neuroticism and workload history on performance. *Personality and Individual Differences*, *36*(2), 447–456. https://doi.org/10.1016/S0191-8869(03)00108-9

Cummings, M. L., Huang, L., Zhu, H., Finkelstein, D., & Wei, R. (2019). The impact of increasing autonomy on training requirements in a UAV supervisory control task. *Journal of Cognitive Engineering and Decision Making*, *13*(4), 295–309. https://doi.org/10.1177/1555343419868917

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development*, *11*(2), 121–136. https://doi.org/10.1080/15248371003699969

Devlin, S. P., Byham, J. K., & Riggs, S. L. (2021). Does what we see shape history? Examining workload history as a function of performance and ambient/focal visual attention. *Transactions on Applied Perception*, *18*(2), 1–17. https://doi.org/10.1145/3449066

Devlin, S. P., Moacdieh, N. M., Wickens, C. D., & Riggs, S. L. (2020). Transitions between low and high levels of mental workload can improve multitasking performance. *IISE Transactions on Occupational Ergonomics and Human Factors*, *8*(2), 72–87. https://doi.org/10.1080/24725838.2020.1770898

Devlin, S. P., & Riggs, S. L. (2018). The effect of video game experience and the ability to handle workload and workload transitions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 736–740. https://doi.org/10.1177/1541931218621167

Edwards, T., Martin, L., Bienert, N., & Mercer, J. (2017). The relationship between workload and performance in air traffic control: Exploring the influence of levels of automation and variation in task demand. *International Symposium on Human Mental Workload: Models and Applications*, 1207–139. https://doi.org/10.1007/978-3-319-61061-0_8

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd ed). Wiley.

Gluckman, J. P., Warm, J. S., Dember, W. N., & Rosa, R. R. (1993). Demand transitions and sustained attention. *The Journal of General Psychology*, *120*(3), 323–337. https://doi.org/10.1080/00221309.1993.9711151

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Wiley-Interscience.

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive energetical framework. *Biological Psychology*, *45*, 73–93. https://doi.org/10.1016/S0301-0511(96)05223-4

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. Routledge.

Jansen, R. J., Sawyer, B. D., van Egmond, R., de Ridder, H., & Hancock, P. A. (2016). Hysteresis in mental workload and task performance: The influence of demand transitions and task prioritization. *Human Factors*, *58*(8), 1143–1157. https://doi.org/10.1177/0018720816669271

Kim, H.-Y. (2015). Statistical notes for clinical researchers: Post-hoc multiple comparisons. *Restorative Dentistry & Endodontics*, *40*(2), 172. https://doi.org/10.5395/rde.2015.40.2.172

Kim, N., House, R., Yun, M. H., & Nam, C. S. (2019). Neural correlates of workload transition in Multitasking: An ACT-R model of hysteresis effect. *Frontiers in Human Neuroscience*, *12*(535), 1–14. https://doi.org/10.3389/fnhum.2018.00535

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(863), 1–12. https://doi.org/10.3389/fpsyg.2013.00863

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Matthews, G., & Desmond, P. A. (2002). Task-induced fatigue states and simulated driving performance. *The Quarterly Journal of Experimental Psychology*, *55A*(2), 659–686. https://doi.org/10.1080/02724980143000505

McKendrick, R., & Harwood, A. (2019). Cognitive workload and workload transitions elicit curvilinear hemodynamics during spatial working memory. *Frontiers in Human Neuroscience*, *13*(405), 1–16. https://doi.org/10.3389/fnhum.2019.00405

Moacdieh, N. M., Devlin, S. P., Jundi, H., & Riggs, S. L. (2020). Effects of workload and workload transitions on attention allocation in a dual-task environment: Evidence from eye tracking metrics. *Journal of Cognitive Engineering and Decision Making*, *14*(2), 132–151. https://doi.org/10.1177/1555343419892184

Morgan, J. F., & Hancock, P. A. (2011). The effect of prior task loading on mental workload: An example of hysteresis in driving. *Human Factors*, *53*(1), 75–86. https://doi.org/10.1177/0018720810393505

Moroney, B. W., Warm, J. S., & Dember, W. N. (1995). Effects of demand transitions on vigilance performance and perceived workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1375–1379. https://doi.org/10.1177%2F154193129503902104

Mracek, D. L., Arsenault, M. L., Day, E. A., Hardy, J. H., & Terry, R. A. (2014). A multilevel approach to relating subjective workload to performance after shifts in task demand. *Human*

Factors, *56*(8), 1401–1413. https://doi.org/10.1177/0018720814533964

Muhs, K. S., Karwowski, W., & Kern, D. (2018). Temporal variability in human performance: A systematic literature review. *International Journal of Industrial Ergonomics*, *64*, 31–50. https://doi.org/10.1016/j.ergon.2017.10.002

O'Rourke, R. (2006, October 25). Unmanned vehicles for U.S. naval forces: Background and issues for congress. *The Navy Department Library*.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., Heisterkamp, S., Willigen, V. B., & Ranke, J. (2021). *Package "nlme."* https://bugs.r-project.org

Rizopoulos, D. (2021). Package 'GLMMadaptive'. *Journal of Computational and Graphical Statistics*, *4*(1). https://doi.org/10.1080/10618600.1995.10474663

RStudio Team (2020). *RStudio: Integrated Development for R.* RStudio, PBC, Boston, MA, http://www.rstudio.com/

Tsukahara, J. S., Harrison, T. L., & Engle, R. W. (2016). The relationship between baseline pupil size and intelligence. *Cognitive Psychology*, *91*, 109–123. https://doi.org/10.1016/j.cogpsych.2016.10.001

Ungar, N. R., Matthews, G., Warm, J. S., Dember, W. N., Thomas, J. K., Finomore, V. S., & Shaw, T. H. (2005). Demand transitions and tracking performance efficiency: Structural and strategic models. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 1523–1526. https://doi.org/10.1177/154193120504901704

CHAPTER 5

Are scan-based eye tracking metrics predictive of workload transition's performance trends over time?

Introduction

It important to understand the operator's visual attention allocation patterns as they manage workload transitions because the operator is often inundated with visual information (Abich et al., 2017; Hobbs & Shivley, 2014; Sibley et al., 2015). However, there is limited research using eye tracking to examine workload transitions (exception Devlin, Byham, & Riggs, 2021; Moacdieh, Devlin, Jundi, & Riggs, 2020) and neither explore whether it can predict performance over time—the focus of this chapter. The goal here is to determine whether visual attention patterns are predictive of the performance trends observed over time during workload transitions. Specifically, we examine several eye tracking metrics as predictors in the growth curve models of performance established in Chapter 4.

Review of previous investigations studying workload transitions over time.

One potential way to explain the different performance trends observed across transition rates in Chapter 4 is to identify and detail the mental processes used during workload transitions. Previous investigations find participants produce a psychophysiological response to workload transitions (Bowers et al., 2014; Boyer et al., 2015; Cerruti et al., 2010; Kim et al., 2019;

McKendrick & Harwood, 2019). There are some limitations with the current investigations, as they do not indicate how a person explicitly used their mental resources to complete the task. As detailed in Chapter 3, *scan-based* eye tracking metrics, i.e., measures capturing the features of visual attention allocation (Poole & Ball, 2006), can quantify and compare visual attention patterns, which has been informative in a variety of settings (as detailed in Chapter 1). Chapter 3 found visual attention patterns clarified multitasking and over time performance trends of workload transitions so of interest is to directly build on those findings by modeling the most informative scan-based metrics as predictors of workload transition performance. Eye tracking has successfully predicted operator needs and mental state in previous research, but to our knowledge, has not attempted to predict workload transition performance (e.g., Barz et al., 2021; Steichen et al., 2013; Ratwani et al., 2010).

Chapter 3 found spread and directness measures were different between constant and transitioning workload, with newer-developed metrics being the most enlightening. For example, stationary gaze entropy, which measures how distributed attention transitions are across the set of AOIs, was lower during workload transitions than during constant workload. Given this was coupled with better multitasking performance, these results suggested visual attention transitions to task-specific areas of the display should not be equal in multitasking environments. These results emphasize the need to include scan-based metrics, especially ones mapped to how visual attention is allocated to each task when studying workload transitions.

Motivation and research questions

The goal of this Chapter 5 is to understand the extent to which scan-based eye tracking metrics can predict performance trends of the three different transition rates (slow, medium, and fast). Specifically, we explored a subset of the scan-based metrics used in Chapter 3 as predictors in the growth curve models of performance established in Chapter 4. For prediction models to be useful in real-world environments, they need to account for the variability amongst individuals, which is a strength of growth curve modeling. Additionally, this modeling approach also allows for a more specific investigation on the predictive capability of scan-based metrics, as they can be specified to predict *baseline* performance and/or how it will *change* over time, allowing for a more informative prediction model (Hoffman, 2015, pp. 286-287).

The scan-based metrics used in this chapter will be the metrics that discriminated between constant and transitioning workload in Chapter 3. We also included at least one metric from the three different aspects of visual attention allocation, i.e., *spread*, *directness*, and *duration* (Moacdieh & Sarter, 2015). Coefficient K was not included in the present investigation because it was found to depend on time, i.e., workload period, which violates an assumption of adding predictors to the model (further explanation provided in the Results section). The list of scan-based metrics used in Chapter 5 is presented in Table 5.1.

Table 5.1 The scan-based metrics explored as predictors for all the previously established response time and accuracy growth curve models

Metric	Definition and calculation
Spread metrics (where are users generally looking?)	
Spatial density	The number of grid cells containing gaze points divided by the total number of cells. A 20×20 evenly-divided grid (128×72 pixels per cell) was created to cover the full screen dimensions. A higher spatial density would indicate a larger dispersion of attention (Goldberg & Kotval, 1999).
Stationary gaze entropy (SGE)	Stationary gaze entropy indicates how equally distributed a person's attention is, with larger values indicating more evenly spread attention across areas of interest (AOI) and lower values indicating more narrowed attention (Krejtz et al., 2015). It is calculated as follows: $$H_s = -\sum_{i \in AOIs} p_i \log_2 p_i$$ where p_i represents the proportion of transitions to the i^{th} state, i.e. the i^{th} AOI (the AOIs are as defined in Figure 2.1) from on all the state transitions based on the Markov property (i.e., transitions to a given state only depend on the current state; Shiferaw et al., 2019). The value is then normalized for comparison (Duchowski, 2017).
Directness metrics (how purposeful are attention transitions?)	
Gaze transition rate [grid cells/s]	The rate of transitions between equal grid cells (Goldberg & Kotval, 1999). A higher rate of transitions indicates lower efficiency. The same grid cells used for spatial density were used here.
Gaze transition entropy (GTE)	The gaze transition entropy represents the randomness and complexity of a person's eye movements, with higher values indicating more randomness and lower efficiency (Krejtz et al., 2015). It is calculated as follows: $$H_t = -\sum_{i \in AOIs} p_i \sum_{j \in AOIs} p_{ij} \log_2 p_{ij}$$ where p_i is as described in stationary gaze entropy, and p_{ij} is the probability of transitioning form state *i* to state *j* in one fixation. Assuming the Markov property holds, this was calculated by counting the number of transitions from *i* to *j* and then dividing by the total number of transitions from *i* (Shiferaw et al., 2018). This was done for each pairing of AOIs (the AOIs are as defined in Figure 2.1). The value is then normalized for comparison (Duchowski, 2017).
Duration metrics (how long, in general, does attention last?)	
Fixation duration [ms]	The amount of time a fixation lasts. A lower mean fixation duration suggests the user is extracting information quickly (Jacob & Karn, 2003).

Previous work has applied growth curve modeling to eye tracking data (e.g., Ayasse & Wingield, 2020; Barr, 2007; Godfroid et al., 2018; Mirman et al., 2008), but rarely as a direct predictor of performance trends over time, especially when workload transitions. Given this investigation is the first of its kind, our research questions are as follows:

1. **RQ 5.1:** Is eye tracking predictive of performance trends over time?
    a. **RQ 5.1a:** Is this a function of workload transition rate?

We expect that multiple scan-based metrics, especially spread and directness metrics, will be predictive of performance trends over time, i.e., multiple metrics will be significant predictors in each growth curve model, but the same scan-based metrics will be significant predictors across the three transition rates per the minimal differences previously observed for these metrics in Chapter 3. If successful, the findings may add the much-needed detail to the theoretical explanations surrounding workload transitions, e.g., *how* mental resources are deployed, and inform display design in complex, multitasking environments, like UAV command and control.

Method

Participants

The same participants from Chapter 4 were used to answer this chapter's research questions. All participants had less than 20% of their raw gaze samples missing as recommended by Komogortsev et al. (2010).

Experimental setup

All the details from the experimental setup of Chapter 4 apply. Participants sat approximately 65 cm from the monitor-mounted Gazepoint HD eye tracker (fs=150 Hz, reported accuracy of 0.5-1˚; Gazepoint, 2019) so their eye movements could be collected.

UAV command and control testbed and tasks

The same UAV command and control testbed from Chapter 4 was used for this chapter's research goals.

Testbed scenarios

The same testbed scenarios from Chapter 4 were relevant to this chapter's research goals.

Procedures

The same procedures from Chapter 4 applied to the present research goals. Participants learned the eye tracker's 9-point calibration procedure during the self-paced informational training session. Participants calibrated their point of gaze to the eye tracker before each transition scenario.

Results

Preprocessing eye tracking data

Raw gaze points, which consist of the positional (x_i, y_i) and temporal information (t_i), were screened for completeness and accuracy via the data quality metric provided by Gazepoint and trials were removed if they did not reach the quality threshold (as previously mentioned). Velocity profiles were then calculated from the raw gaze points (x_i, y_i, t_i) by differentiating with a six-tap Savitzky-Golay filter of degree 2 (Krejtz et al., 2016). An I-VDT event detection algorithm (Komogortsev & Karpov, 2013) was used to determine fixations due to the increased sampling rate. The velocity threshold for fixations was determined by the adaptive algorithm outlined in Nyström & Holmqvist (2010), and it ranged 25.6-60.8 degrees/s across all trials. Then, individual fixations were determined as clusters of raw gaze points that were below the trial's velocity threshold, a maximum of 110 pixels from each other (i.e., ~1° visual angle), and occurred within a minimum of 80 ms from each other.

Model fitting process to determine a conditional growth curve model

The ultimate goal of this modeling process is to identify the eye tracking metric(s) that predict performance trends over time when workload transitions. The five scan-based metrics from Table 5.1 were explored as time-invariant predictors (TIP) for each growth curve model. Response time and accuracy were modeled in RStudio (version 1.3.1093; RStudio Team, 2020) and as separate growth curve models for each transition rate (with packages *lme4* and *GLMMadaptive*, respectively; Bates et al., 2015; Rizopoulous, 2021), but response time for fast transitions was excluded due to no presence of individual differences. A time-invariant predictor is a measure of the individual that is not expected to change over time or can only be reliably measured once in the experiment (Hoffman, 2015, p. 312). In the present research, the eye tracking metrics were modeled as time-invariant predictors because they served as measures of the individual's task completion strategy and were only calculated once per person per testbed scenario. Exploring the predictive ability of these scan-based metrics consisted of following the two-phase modeling process recommended by (Hoffman, 2021):

1. The first phase was to determine the bivariate relationship between each eye tracking metric and each growth curve model. This consisted of comparing model fits of each unconditional growth curve model, i.e., the best fitting growth curve model where time is the only predictor variable, e.g., Equation (4.5) to a conditional growth curve model, i.e., a growth curve model that includes more than just predictors of time, with a single scan-based metric. There were two types of conditional growth curve models that were first considered: (1) one where the scan-based metric was an additive effect, i.e., an additive time-invariant predictor, and (2) one where the scan-based metric was a cross-level interaction effect with the linear time slope of the unconditional growth curve model, i.e., a cross-level time-invariant predictor. An additive time-invariant predictor predicts baseline performance, i.e., the model's intercept, where a cross-level time-invariant predictor predicts how performance changes over time, i.e., the model's linear time slope. Equation 5.1 shows an example of the scan-based metric included as an additive time-invariant predictor (γ_{01}) in a fixed cubic, random intercept growth curve model for each person (i) over time (t). Equation 5.2 shows an example of the scan-based metric as a cross-level time-invariant predictor (γ_{11}) in a fixed cubic, random intercept growth curve model.

$$y_{ti} = \beta_{0i} + \beta_1(Time_{ti}) + \beta_2(Time_{ti})^2 + \beta_3(Time_{ti})^3 \tag{5.1}$$

$$where,\ \beta_{0i} = \gamma_{00} + \gamma_{01}(Scan - based\ metric_i) + U_{0i},\ \beta_1 = \gamma_{10},\ \beta_2 = \gamma_{20}, \beta_3 = \gamma_{30}$$

$$y_{ti} = \beta_{0i} + \beta_1(Time_{ti}) + \beta_2(Time_{ti})^2 + \beta_3(Time_{ti})^3 \tag{5.2}$$

$$where,\ \beta_{0i} = \gamma_{00} + \gamma_{01}(Scan - based\ metric_i) + U_{0i},$$
$$\beta_1 = \gamma_{10} + \gamma_{11}(Scan - based\ metric_i),\ \beta_2 = \gamma_{20},\ \beta_3 = \gamma_{30}$$

In summary, establishing the bivariate relationship between each scan-based metric and performance growth curve model separately better specifies the predictive capability the scan-based metric has on the observed performance trends. Likelihood ratio tests assessed model fit where significance for an additive time-invariant predictor was set at α=0.05 and significance for a cross-level time-invariant predictor was set at α=0.10 (Mathieu et al., 2012).

2. The second phase of the analysis began once the bivariate relationship between all unconditional and conditional growth curve models were established. The goal of this phase was to answer the research questions by determining: (a) the predictive capability of an eye tracking metric when it was combined with other significant eye tracking metrics and (b) differences in combined conditional growth curve model across transition rates. Each scan-based metric that had a significant bivariate relationship was strategically combined into a single model, specifically by building upon bivariate models with a cross-level time-invariant predictor and then including additive time-invariant predictors thereafter. Likelihood ratio tests were used to determine model fit across conditional growth curve models (additive: α=0.05 and cross-level: α=0.10; Mathieu et al., 2012), but when that failed to distinguish a better

model fit, Bayesian Information Criterion (BIC) determined the final conditional growth curve model (Hoffman, 2015, p. 271). An example of when this happened was when the scan-based metrics in the models were the same, but they differed in the type of time-invariant predictor, e.g., additive vs. cross-level. Psuedo-R^2 was calculated as the effect size measure for each time-invariant predictor in the combined model (Raudenbush & Bryk, 2002; Singer & Willet, 2003). The results from the first phase, i.e., the bivariate relations, are briefly mentioned in the text, but the results from the second phase, i.e., all the estimates and fit statistics of each conditional growth curve model are presented in detail with the final combined model interpreted in the text.

For slow transitions, fixation duration and stationary gaze entropy predict the baseline response time and its trend over time

The unconditional growth curve model of response time during slow transitions had a positive fixed cubic time slope and a random intercept ($\chi^2(1)= 63.232$, $p<.001$; model fitting details in Table 4.1). The first phase of the analysis found the following eye tracking metrics had a significant bivariate relation: stationary gaze entropy (additive and cross-level), gaze transition entropy (cross-level: $\chi^2(1)= 3.592$, $p=.058$), average fixation duration (additive and cross-level), and gaze transition rate (additive: $\chi^2(1)= 4.712$, $p=.030$). In the second phase of the analysis, stationary gaze entropy and average fixation duration remained significant, but the BIC was smaller when stationary gaze entropy was a cross-level time-invariant predictor and average fixation duration was an additive time-invariant predictor ($\chi^2(1)= 4.490$, $p=.034$). Therefore, this was determined as the final conditional growth curve model, i.e., Model 5 in Table 5.2.

Specifically, longer average fixation duration (*M*=125.2 ms, *SD*=17.3 ms) predicted faster baseline response time during slow transitions (Model 5's rescaled γ_{01}=-0.005, 95% CI: [-0.001, -0.00002]). In other words, for every one standard deviation increase in average fixation duration, and assuming all other predictors remain equal, response time improved, i.e., sped up, by 0.09 s. Stationary gaze entropy was a significant cross-level time-invariant predictor, meaning it predicted *changes in* response time during slow transitions. In general, it found larger stationary gaze entropy (*M*=0.63, *SD*=0.08) predicted more severe decrements in response time during slow transitions (Model 5's rescaled γ_{11}= -0.03, 95% CI: [-0.0002, -0.003]). Specifically, for every one standard deviation increase in stationary gaze entropy, the estimated total decrement in response time would be 0.2 s longer. Table 5.2 shows the results of all the combined models and Figure 5.1 depicts how response time follows a cubic trend over time, but the scan-based metrics predict baseline response time and how it changes over time.

Table 5.2 Slow transitions scenario's scaled estimates of significant scan-based metrics as time-invariant predictors in the combined conditional growth curve models of primary task response time (time-invariant predictors=TIP; AFD=average fixation duration; SGE=stationary gaze entropy). Final model is bolded and highlighted in gray

	Unconditional growth curve model (fixed positive cubic time slope, random intercept)	Model 1 (conditional growth curve model with SGE as an additive TIP)	Model 2 (conditional growth curve model with SGE as a cross-level TIP)	Model 3 (conditional growth curve model with AFD as an additive TIP)	Model 4 (conditional growth curve model with AFD as a cross-level TIP)	**Model 5** (model 2 with AFD as an additive TIP)	Model 6 (model 4 with SGE as an additive TIP)
Intercept (γ_{00})	3.247*** (0.044)	3.244*** (0.040)	3.244*** (0.040)	3.245*** (0.040)	3.245*** (0.039)	3.244*** (0.037)	3.244*** (0.037)
Fixed linear time slope (γ_{10})	2.482*** (0.271)	2.484*** (0.271)	2.002*** (0.344)	2.475*** (0.271)	2.841*** (0.337)	1.994*** (0.344)	2.846*** (0.337)
Fixed quadratic time slope (γ_{20})	-5.695*** (0.658)	-5.703*** (0.658)	-5.685*** (0.657)	-5.678*** (0.658)	-5.689*** (0.657)	-5.669*** (0.657)	-5.697*** (0.657)
Fixed cubic time slope (γ_{30})	3.315*** (0.414)	3.320*** (0.414)	3.308*** (0.415)	3.304*** (0.414)	3.309*** (0.414)	3.298*** (0.414)	3.315*** (0.414)
Fixed additive TIP (γ_{01})	-	0.108* (0.040)	-0.002 (0.063)	-0.113** (0.040)	-0.025*** (0.063)	AFD: -0.087* (0.040) SGE: -0.031 (0.062)	AFD: 0.002 (0.062) SGE: 0.081* (0.039)
Fixed cross-level TIP (γ_{11})	-	-	0.494* (0.216)	-	-0.374† (0.204)	0.496* (0.216)	-0.375† (0.204)
Intercept variance ($\tau_{U_0}^2$)	0.038	0.027	0.027	0.026	0.026	0.020	0.020
Residual variance (σ_e^2)	1.739	1.731	1.735	1.739	1.737	1.735	1.737

Psuedo-R^2 for ($\tau^2_{U_0}$)	-	0.299	0.298	0.333	0.331	0.472	0.472
Psuedo-R^2 for (σ^2_e)	-	-	0.002	-	0.001	0.002	0.002
Fit Statistics							
LL	-	-3942.5	-3939.9	-3942.3	-3940.6	-3937.7	-3938.6
χ^2 df	-	1	1	1	1	1	1
χ^2	-	6.719	5.201	7.161	3.339	4.490	4.031
p	-	0.010**	0.023*	0.007**	0.068†	0.034*	0.045*
BIC	-	-	-	-	-	7945.1	7947.0

Significance codes: '***' indicates p<0.001, '**' indicates p<0.01, '*' indicates p<0.05, '†' indicates p<0.10

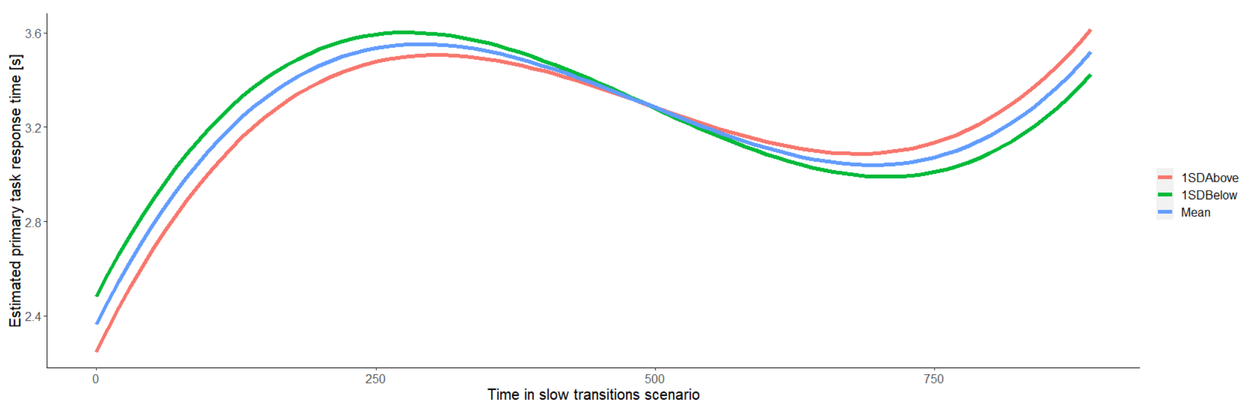


Figure 5.1 Slow transitions' combined conditional growth curve models for primary task response time, when each scan-based metric is at its mean and ±1 standard deviation (SD) from it

For medium transitions, stationary gaze entropy predicts baseline response time

The unconditional growth curve model of response time during the medium transitions scenario had a positive quadratic fixed effect of time and a random intercept (χ^2(1)=7.263, p=.007; model fitting details in Table 4.2). Stationary gaze entropy was the only metric that had a significant bivariate relation with the unconditional growth curve model, specifically as an additive time-invariant predictor ($\chi^2(1)$ =7.986, p=.005), making it the final conditional growth curve model by default, i.e., Model 1 in Table 5.3.

Specifically, larger stationary gaze entropy (M=0.60, SD=0.09) predicted slower baseline response time during medium transitions (Model 1's rescaled γ_{01}=1.27, 95% CI: [0.42, 2.13]). In other words, for every one standard deviation increase in stationary gaze entropy, baseline response time slowed by 0.11 s. Table 5.3 shows the results of all the combined models and Figure 5.2 shows how primary task response time would follow the predicted quadratic trend, but stationary gaze entropy predicted baseline response time speeds.

Table 5.3 Medium transitions scenario's scaled estimates of significant scan-based metrics as time-invariant predictors in the combined conditional growth curve models of primary task response time (time-invariant predictors=TIP; SGE=stationary gaze entropy). Final model is bolded and highlighted in gray

	Unconditional growth curve model (fixed positive quadratic time slope, random intercept)	**Model 1** (Conditional growth curve model with SGE as an additive TIP)	Model 2 (Conditional growth curve model with SGE as a cross-level TIP)
Intercept (γ_{00})	3.850*** (0.043)	3.850*** (0.039)	3.850*** (0.039)
Fixed linear time slope (γ_{10})	-0.257* (0.107)	-0.255* (0.107)	-0.192 (0.207)
Fixed quadratic time slope (γ_{20})	0.289** (0.107)	0.288** (0.107)	0.288** (0.107)
Fixed additive TIP (γ_{01})	-	0.116** (0.039)	0.132* (0.060)
Fixed cross-level TIP (γ_{11})	-	-	-0.066 (0.184)
Intercept variance ($\tau_{U_0}^2$)	0.040	0.026	0.026
Residual variance (σ_e^2)	1.880	1.880	1.879
Psuedo-R^2 for $\tau_{U_0}^2$	-	0.330	0.330
Psuedo-R^2 for σ_e^2	-	-	0.0001
Fit Statistics			
LL	-4611.5	-4607.5	-4607.4
χ^2 df	-	1	1
χ^2	-	7.986	0.129
p	-	0.005**	0.718

Significance codes: '***' indicates p<0.001, '**' indicates p<0.01, '*' indicates p<0.05 '*', '†' indicates p<0.10

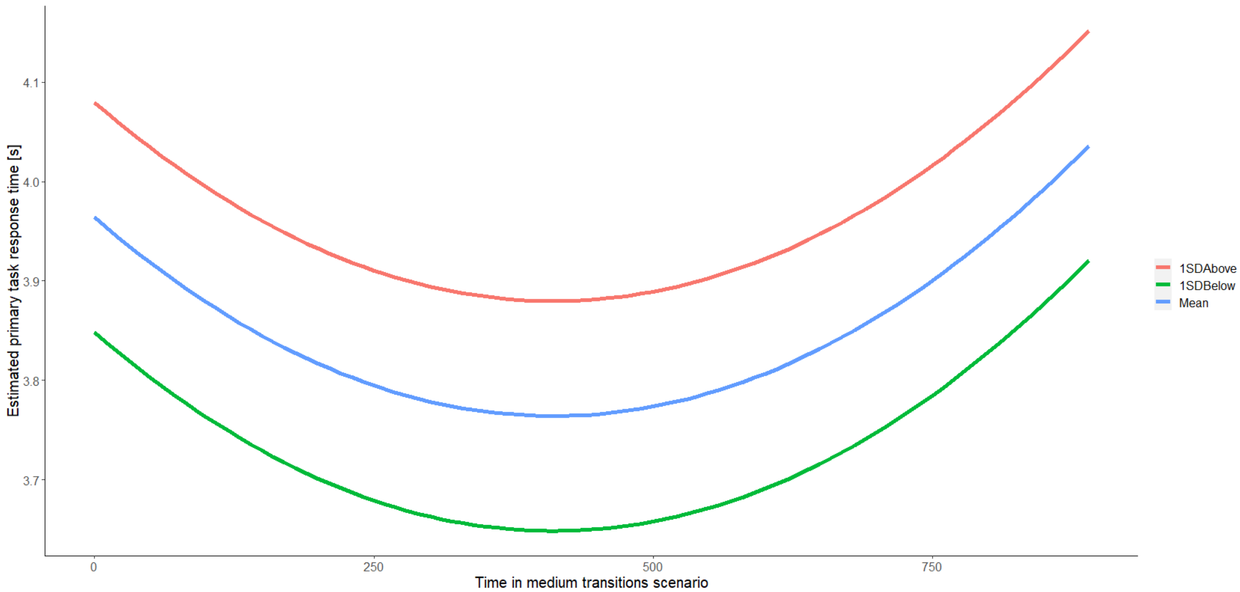


Figure 5.2 Medium transitions' combined conditional growth curve model of primary task response time, when stationary gaze entropy is at its mean and ±1 standard deviation from it

For slow transitions, stationary gaze entropy predicts baseline primary task accuracy

The unconditional growth curve model of accuracy during slow transitions had a negative fixed cubic time slope and a random intercept ($\chi^2(1)$=66.58, p<.001; model fitting details in Table 4.4). Stationary gaze entropy was the only significant time-invariant predictor in the bivariate investigation, specifically as an additive time-invariant predictor ($\chi^2(1)$=7.84, p<.005), making it the final conditional growth curve model, i.e., Model 1 in Table 5.4.

The model found larger stationary gaze entropy values predicted worse baseline accuracy during slow transitions (Model 1's rescaled γ_{01}= -2.70, 95% CI: [-4.47, -0.92]). For every one standard deviation increase in stationary gaze entropy (M=0.63, SD=0.08) estimated baseline accuracy to worsen by 5.0%. Table 5.4 shows the results of all the combined models and Figure 5.3 shows how accuracy follows a cubic trend over time, but stationary gaze entropy predicts baseline accuracy rates.

Table 5.4 Slow transitions scenario's scaled estimates of significant scan-based metrics as time-invariant predictors in the combined conditional growth curve models of primary task accuracy (time-invariant predictors=TIP; SGE=stationary gaze entropy). Final model is bolded and highlighted in gray

	Unconditional growth curve model (fixed negative cubic time slope, random intercept)	**Model 1** (Conditional growth curve model with SGE as an additive TIP)	Model 2 (Conditional growth curve model with SGE as a cross-level TIP)
Intercept (γ_{00})	0.776*** (0.088)	0.777*** (0.078)	0.777*** (0.078)
Fixed linear time slope (γ_{10})	-2.992*** (0.273)	-2.992*** (0.273)	-2.999*** (0.305)
Fixed quadratic time slope (γ_{20})	6.829*** (0.608)	6.829*** (0.608)	6.829*** (0.608)
Fixed cubic time slope (γ_{30})	-4.068*** (0.367)	-4.067*** (0.367)	-4.068*** (0.367)
Fixed additive TIP (γ_{01})	-	-0.226*** (0.078)	-0.227** (0.087)
Fixed cross-level TIP (γ_{11})	-	-	0.0061 (0.180)
Intercept variance ($\tau^2_{U_0}$)	0.206	0.156	0.156
Psuedo-R^2 for $\tau^2_{U_0}$	-	0.244	0.244
Fit Statistics			
LL	-2352.8	-2348.9	-2348.9
χ^2 df	-	1	1
χ^2	-	7.840	0.001
p	-	0.005**	0.980

Significance codes: '***' indicates $p<0.001$, '**' indicates $p<0.01$, '*' indicates $p<0.05$ '*', '†' indicates $p<0.10$

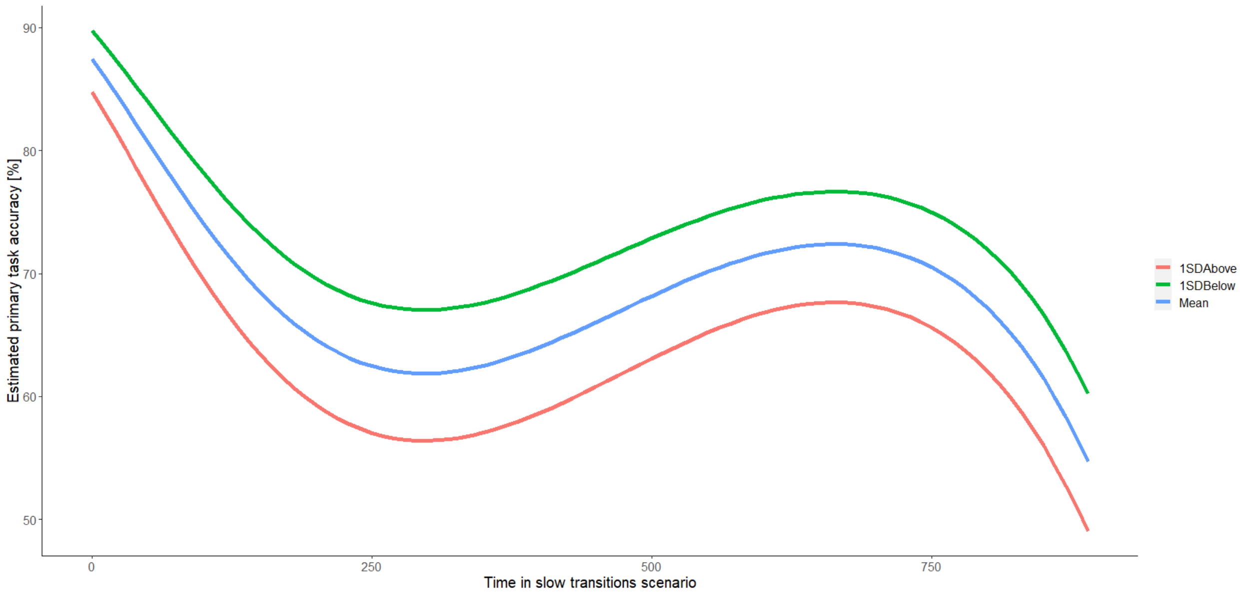


Figure 5.3 Slow transitions' combined conditional growth curve model of primary task accuracy, when stationary gaze entropy is at its mean and ±1 standard deviation (SD) from it

For medium transitions, stationary gaze entropy predicts baseline primary task accuracy

The unconditional growth curve model of accuracy during the medium transitions scenario had a negative fixed cubic time slope, a random intercept, and random linear time slope ($\chi^2(1)$=2.31, p=0.09; model fitting details in Table 4.5). In the first phase of the analysis, both stationary gaze entropy and average fixation duration were significant additive time-invariant predictors. However, average fixation duration was no longer significant when combining both scan-based metrics as additive time-invariant predictors into a single model, making the final conditional growth curve model to include stationary gaze entropy as an additive time-invariant predictor, i.e., Model 1 in Table 5.5.

Specifically, <u>larger</u> stationary gaze entropy predicted <u>worst</u> baseline accuracy during medium transitions (i.e., Model 1's γ_{01}=-2.35, 95% CI: [-3.60, -1.09]). For example, every one standard deviation <u>increase</u> in stationary gaze entropy (M=0.60, SD=0.09), baseline accuracy would decrease by 4.8%. Table 5.5 shows the results of all the combined models and Figure 5.4 shows how accuracy follows a cubic trend over time, but stationary gaze entropy predicts its baseline accuracy rates.

Table 5.5 Medium transitions scenario's scaled estimates of significant scan-based metrics as time-invariant predictors in the combined conditional growth curve models of primary task accuracy (time-invariant predictors=TIP; AFD=average fixation duration; SGE=stationary gaze entropy). Final model is bolded and highlighted in gray

	Unconditional growth curve model (fixed negative cubic time slope, random intercept and linear time slope)	**Model 1** (Conditional growth curve model with SGE as an additive TIP)	Model 2 (Conditional growth curve model with SGE as a cross-level TIP)	Model 3 (Conditional growth curve model with AFD as an additive TIP)	Model 4 (Conditional growth curve model with AFD as a cross-level TIP)	Model 5 (Model 1 and Model 3 combined)
Intercept (γ_{00})	0.734*** (0.070)	0.735*** (0.058)	0.735*** (0.058)	0.735*** (0.064)	0.735*** (0.064)	0.735*** (0.057)
Fixed linear time slope (γ_{10})	-0.387 (0.240)	-0.389 (0.239)	-0.178 (0.298)	-0.387 (0.240)	-0.526 (0.329)	-0.389 (0.239)
Fixed quadratic time slope (γ_{20})	1.334* (0.5950)	1.337* (0.5927)	1.336* (0.5934)	1.333* (0.5947)	1.334* (0.595)	1.336* (0.593)
Fixed cubic time slope (γ_{30})	-0.826* (0.381)	-0.828* (0.379)	-0.827* (0.380)	-0.826* (0.380)	-0.826* (0.380)	-0.828* (0.380)
Additive TIP (γ_{01})	-	-0.213*** (0.056)	-0.173** (0.067)	0.154** (0.057)	0.132* (0.061)	AFD: 0.057 (0.065) SGE: -0.182** (0.061)
Cross-level TIP (γ_{11})	-	-	-0.218 (0.220)	-	0.145 (0.235)	-
Intercept variance ($\tau^2_{U_0}$)	0.147	0.10	0.095	0.120	0.120	0.09
Linear slope variance ($\tau^2_{U_1}$)	0.009	0.009	0.008	0.009	0.009	0.009
Psuedo-R^2 for $\tau^2_{U_0}$	-	0.351	0.351 0.107	0.180	0.179 0.029	0.365

and $\tau^2_{U_1}$						
Fit Statistics						
LL	-2991.9	-2986.3	-2985.7	-2989.2	-2988.9	-2985.9
χ^2 df	-	1	1	1	1	Model 1: 1 Model 3: 1
χ^2	-	11.37	1.09	5.57	0.41	Model 1: 0.71 Model 3: 6.52
p	-	<0.001	0.296	0.018*	0.524	Model 1: 0.398 Model 3: 0.011*

Significance codes: '***' indicates p<0.001, '**' indicates p<0.01, '*' indicates p<0.05 '*', '†' indicates p<0.10

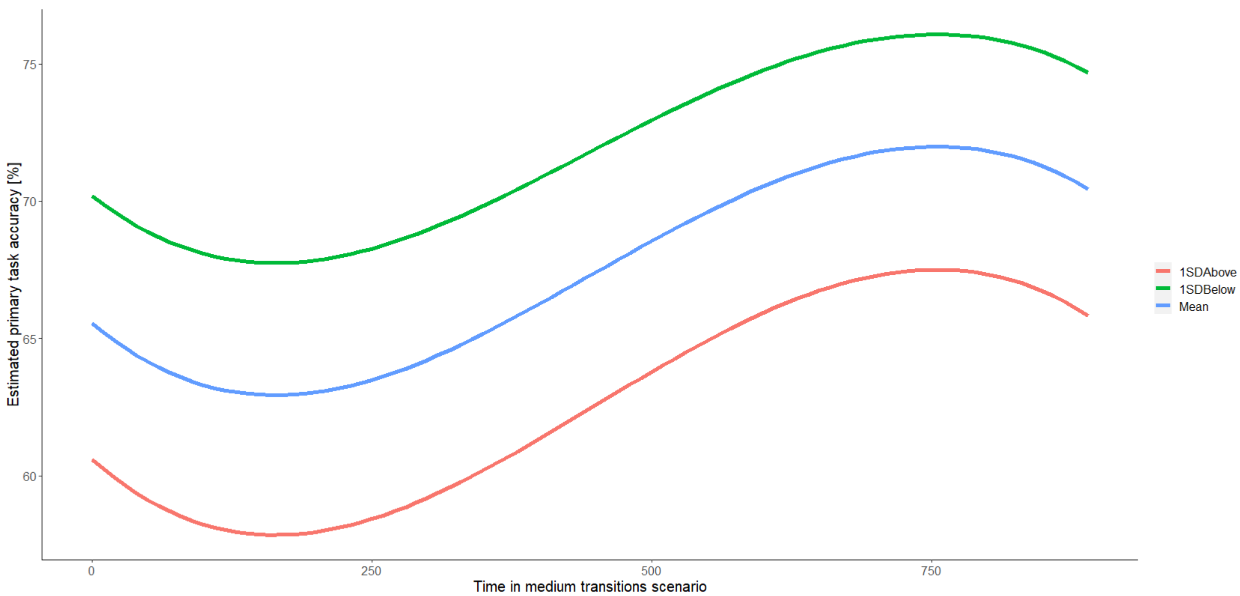


Figure 5.4 Medium transitions' combined conditional growth curve model of primary task accuracy, when stationary gaze entropy is at its mean and ±1 standard deviation (SD) from it

For fast transitions, stationary gaze entropy and gaze transition rate predict baseline accuracy and its change over time

The unconditional growth curve model had a negative fixed cubic time slope and a random intercept ($\chi^2(1)=16.27$, $p<.001$; model fitting details in Table 4.6). In the first phase of the analysis, stationary gaze entropy was a significant additive time-invariant predictor and gaze transition rate was a significant cross-level time-invariant predictors. Model fit significantly improved when both time-invariant predictors were combined in to a single model, making it the final conditional growth curve model (Model 5 in Table 5.6).

Specifically, larger stationary gaze entropy predicted worse baseline accuracy during the fast transitions scenario (Model 5's rescaled γ_{01}= -2.31, 95% CI: [-4.33, -0.44]). Specifically, for every one standard deviation increase in stationary gaze entropy (*M*=0.61, *SD*=0.08), and assuming all other predictors are held equal, baseline accuracy was lower by 4.2%. As for gaze transition rate, it predicted that larger gaze transition rates would lead to a less severe accuracy decrement during the fast transitions (Model 5's rescaled γ_{11}=0.003, 95% CI: [-0.00007, 0.006]). Assuming all other predictors are held equal, for every standard deviation above the mean gaze transition rate (*M*=2.7 grids/s, *SD*=0.7), accuracy was estimated to have a decrement that was 3.9% less than a person with a mean gaze transition rate. Table 5.6 shows the results of all the combined models and Figure 5.5 shows how accuracy would follow the cubic trend over time, but a larger stationary gaze entropy and a smaller gaze transition rate would predict worst initial accuracy and for it to get even worst during fast transitions.

Table 5.6 Fast transitions scenario's scaled estimates of significant scan-based metrics as time-invariant predictors in the combined conditional growth curve models for primary task accuracy (time-invariant predictors=TIP; GTR=gaze transition rate; SGE=stationary gaze entropy). Final model is bolded and highlighted in gray.

	Unconditional growth curve model (fixed negative cubic time slope and random intercept)	Model 1 (Conditional growth curve model with SGE as an additive TIP)	Model 2 (Conditional growth curve model with SGE as a cross-level TIP)	Model 3 (Conditional growth curve model with GTR as an additive TIP)	Model 4 (Conditional growth curve model with GTR as a cross-level TIP)	**Model 5** (Model 4 with SGE as an additive TIP)
Intercept (γ_{00})	0.598*** (0.081)	0.597*** (0.075)	0.598*** (0.075)	0.597*** (0.081)	0.598*** (0.081)	0.598*** (0.075)
Fixed linear time slope (γ_{10})	-1.127*** (0.229)	-1.127*** (0.229)	-0.830** (0.293)	-1.127*** (0.229)	-1.332*** (0.269)	-1.333*** (0.269)
Fixed quadratic time slope (γ_{20})	2.805*** (0.495)	2.806*** (0.495)	2.808*** (0.494)	2.805*** (0.495)	2.809*** (0.495)	2.810*** (0.495)
Fixed cubic time slope (γ_{30})	-1.772*** (0.278)	-1.772*** (0.278)	-1.775*** (0.278)	-1.772*** (0.278)	-1.773*** (0.278)	-1.773*** (0.278)
Fixed additive TIP (γ_{01})	-	-0.172* (0.083)	-0.104 (0.110)	0.020 (0.072)	-0.076 (0.094)	SGE: -0.182* (0.086) GTR: -0.130 (0.092)
Fixed cross-level TIP (γ_{11})	-	-	-0.306 (0.200)	-	0.229† (0.134)	0.229† (0.134)
Intercept variance ($\tau^2_{U_0}$)	0.172	0.142	0.142	0.171	0.172	0.141
Psuedo-R^2 for $\tau^2_{U_0}$	-	0.175	0.174	0.002	0.001	0.182
Fit Statistics						
LL	-2422.2	-2419.8	-2418.8	-2422.1	-2420.3	-2417.8
χ^2 df	-	1	1	1	1	1

χ^2	-	4.78	1.86	0.06	3.67	4.97
p	-	0.029*	0.173	0.805	0.056†	0.027*

Significance codes: '***' indicates p<0.001, '**' indicates p<0.01, '*' indicates p<0.05, '†' indicates p<0.10

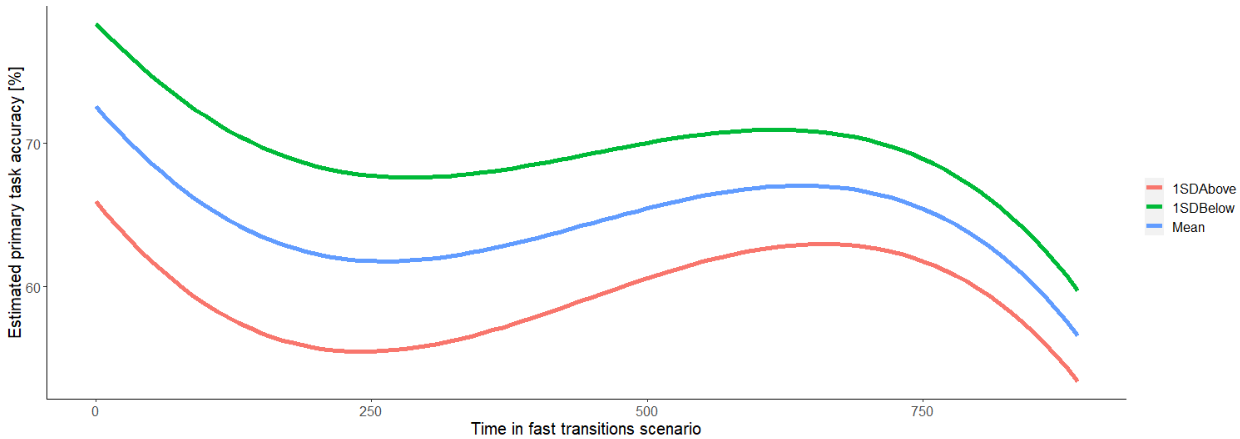


Figure 5.5 Fast transitions' combined conditional growth curve model of primary task accuracy, when each scan-based metric is at its mean and ±1 standard deviation (SD) from it

Table 5.7 summarizes the present results by identifying and interpreting the scan-based metric(s) that were significant time-invariant predictors in the final conditional growth curve models for each performance metric in each testbed scenario.

Table 5.7 Summary of the significant time-invariant predictors. A ✓ means the scan-based metrics was significant in the final combined model. The last column describes the takeaways from each of the final combined models

Transition rate	Performance metric	*Spread metric*		*Directness metric*		*Duration metric*	Takeaway
		Spatial density	*Stationary gaze entropy*	*Gaze transition rate*	*Gaze transition entropy*	*Average fixation duration*	
SLOW	*Response time*		✓			✓	• Longer *average fixation duration* predicts faster baseline response times • Larger *stationary gaze entropy* predicts larger decrements in response time over time
	Accuracy		✓				• Larger *stationary gaze entropy* predicts lower baseline accuracy rates
MEDIUM	*Response time*		✓				• Larger *stationary gaze entropy* predicts slower baseline response times
	Accuracy		✓				• Larger *stationary gaze entropy* predicts lower baseline accuracy rates
FAST	*Accuracy*		✓	✓			• Larger *stationary gaze entropy* predicts lower baseline accuracy • Larger *gaze transition rate* predicts smaller decrements in accuracy over time

Discussion

The goal of this research was to assess whether scan-based metrics are predictive of performance trends (RQ 5.1) and if it depended on workload transition rate (RQ 5.1a). Here, we found these metrics had predictive capability and it was a function of the performance measure, i.e., response time and accuracy, and workload transition rate. The final combined model of slow transitions found the average duration of visual attention (fixation duration) and how dispersed its transitions are across the AOIs (stationary gaze entropy) are predictive of both baseline performance and/or its trends over time. For medium transitions, the dispersion of visual attention transitions across the AOIs (stationary gaze entropy) was predictive of baseline performance. For fast transitions, the dispersion of visual attention transitions across the AOIs (stationary gaze entropy) is predictive of baseline accuracy whereas the average pace of visual attention changes in general (gaze transition rate) is predictive of performance trends over time. Our expectations were mostly met, but with caveats. For instance, across all transition rates, stationary gaze entropy was predictive of baseline performance and/or it trend over time, which is consistent with other work that studies scan-based metrics in realistic environments (Shiferaw et al., 2019). However, the same metrics were not always predictive for each growth curve model, further supporting the notion that workload transition rate matters (Chapter 4).

Implications of stationary gaze entropy being a significant predictor across all transition rates

The ability of stationary gaze entropy to predict performance may lie in the inclusion of context-driven AOIs and the Markov property—i.e., *memoryless* transitions. The Markov property is true if, "the probability distribution of future states of the process conditioned on both the past and present states depends only on the present state" (Gudivada et al., 2015). For example, transitions to AOIs on the display is only dependent on the current AOI the participant is looking at. Here the long-term probabilities are the proportion of transitions that go to each AOI (Shiferaw et al., 2019). Stationary gaze entropy may have high predictive capability because it is a measuring spread based on *active, directionally specific* transitions between AOIs (Shic et al., 2008) and the certainty of those transitions. Thus, providing a single quantitative value on the dynamics of visual attention transitions across AOIs, given both micro (i.e., where current visual attention is transitioning to) and macro (i.e., how the proportion of visual attention to each AOI compare to each other over time) visual attention allocation patterns. Additionally, the current success of assuming transitions between AOIs are a memoryless process, i.e., the Markov property, shows practical potential in deploying effective adaptive assistance based on in visual attention allocation. The eye tracker can intermittently lose contact with the operator's point of gaze, so it is imperative adaptive assistance can provide accurate predictions without requiring the entire scanpath. Vetting this approach is essential as its applicability may change depending on AOI definition or task paradigm.

The overwhelming predictive capability of stationary gaze entropy may not only be due to the Markov property as gaze transition entropy, which also relies on the Markov property, was not nearly as effective in predicting performance. This is actually consistent with some previous work that studied both entropy metrics in a simulated driving environment (e.g., Shiferaw et al., 2018). The predictive capability of stationary gaze entropy may also be due to it being a spread metric, i.e., measuring where someone looks, as gaze transition entropy is a directness metric, i.e., measuring the efficiency of someone's scan. However, between the two spread metrics, i.e., spatial density and stationary gaze entropy, only stationary gaze entropy predicted performance trends, meaning the way spread metrics are defined matters. Spatial density may have been and ineffective predictor because it reduces the display into a uniform grid, providing no context on how each grid cell relates to the task or how often it is viewed. Oppositely, the AOIs of stationary gaze entropy had a direct mapping with a testbed task, which inherently provides more semantic information than a grid cell. Specifically, it informs how participants relied on their visual attention to manage tasks as workload transitioned. Simply knowing how much of the display was viewed at one point in time, which is essentially what spatial density measures, could be more dependent of display design than the participant's visual attention allocation patterns (Moacdieh & Sarter, 2015), potentially making it uninformative on the management of workload transitions. In summary, where *and* how frequent visual attention transitioned to *task-relevant areas* is a better indicator of performance than where it landed in general.

Finally, stationary gaze entropy was not only predictive of performance, but also informs display design. Stationary gaze entropy suggests having a balanced number of transitions between the AOIs results in worse performance across all transition rates. Given this study had a primary task, we conjecture that performance improved when most of the transitions were to the primary task's AOI (i.e., Video Feed panel). From a design perspective, attention should be directed to a primary task and minimized elsewhere either via design features and/or external assistance. When reviewing suggested layouts of current UAV command and control tasks, we found that AOIs associated with each task are typically dispersed across the display and organized into several subgroups, just like this testbed had (Feitshans et al., 2008; Foroughi et al., 2019). It may be strategic to reorganize these displays based on the operator's priorities, especially if multitasking between tasks is not equal.

Implications on theory: Effort regulation may manifest differently depending on sensory modality and depend on the features of the transition and environment

The interpretations from the final combined models add to the existing theory on workload transitions. To date, there has been limited work that expands upon the two explanations—i.e., effort regulation and resource depletion—even when psychophysiological measures are included (e.g., Bowers et al., 2014; Kim et al., 2019). Here, the current findings, allow us a unique opportunity to build upon the effort regulation explanation, given it stipulates that workload transition performance is a function of how mental resources are relied upon. Specifically, it states:

> *Workload transition performance is dependent on the individual actively* ***appraising****, recruiting, and* ***deploying*** *the requisite amount of* ***mental resources*** *for the* ***present workload****. Performance is stable as long as the appraisal is correct and workload does not reach levels of overload (Hockey, 1997).*

We expand upon the bolded terms in the definition above. Specifically, we will add specifications to the theory regarding: (a) *what* in the environment is appraised, and (b) *how* mental resources are deployed, (c) the *type* of mental resource. The only aspect of this explanation we do not address is resource recruitment, but future work ideally can (see alternatives presented in Conclusion).

First, we propose workload transition performance is not only dependent on appraising the amount of workload, but it is also dependent on appraising the rate in which workload transitions. The different significant predictors across the conditional growth curve models support this notion. Although all models predicted performance to worsen when visual attention transitions were more evenly distributed across tasks (i.e., when stationary gaze entropy was larger), there is some evidence different visual attention strategies across the three transition rates improved performance. For example, for slow transitions, longer periods of visual attention predicted improved baseline response time, meaning performing quickly during slow transitions required more thoughtful cognitive processing (Holmqvist et al., 2011; Poole & Ball, 2006). This may be why larger performance differences were observed between periods of low and high workload during slow transitions in Chapter 4, as thoughtful cognitive processing was more limited during high workload. For fast transitions, frequent general attention shifts predicted improved baseline accuracy and a smaller decrement over time suggesting performing well during fast transitions requires more frequent changes in attention (Yang et al., 2018). Chapter 4 finds performance was quicker during fast transitions, potentially because the more frequent change in general attention may lend to detecting targets faster. The significance of stationary gaze entropy during medium transitions also explains why individual differences were detected for both baseline and over time performance in Chapter 4—it was more dependent on the task strategy which stationary gaze entropy holistically captures. Our results show performance improves when visual attention patterns account for transition rate.

Second, we propose the deployment of resources should consider the way in which visual attention is allocated. Namely, its location, efficiency, and time span. This is evident by one scan-based eye tracking metric from each category—i.e., *spread*, *directness*, and *duration*—being a significant predictor for at least one performance trend over time. We propose the deployment of mental resources needs to consider the questions regarding the scan-based metric categories: *where are people generally looking?* (spread), *how efficiently are people looking?*

(directness), and *how long are people looking?* (duration; Moacdieh & Sarter, 2015).

Third, we propose amending the effort regulation explanation to specify that workload transition performance is dependent on the type of mental resource. Here we propose the effort regulation explanation should say workload transition performance hinges on having the requisite amount of *visual* attentional resources. This is supported by the fact that at least one scan-based metric was a significant predictor of performance. Considering the entire experiment was largely visual in nature, this finding is not surprising; however, this highlights how the applicability of this theory may hinge on what sensory modality is considered. This supports the premise of the Multiple Resource Model, which posits different sensory modalities draw from separate attentional resources (Wickens, 1980). Future work should examine whether this holds true for other modalities (i.e., auditory, tactile, etc.). We believe detailing the effort regulation explanation in this way leads to a better understanding and expectation of how people perform and allocate their visual attention during workload transitions.

Implication for design: Technology design for each transition rate based on scan-based metrics

Workload transitions continue to be an important feature of the environment that needs to be taken into consideration when designing visual displays, but currently there is limited design guidance that exists. Beyond prediction, the current results also provide new information on the success of task strategy. Therefore, design guidance for each workload transition rate is outlined in Table 5.8.

Table 5.8 Findings, design guidance, and design example across all transition rates and for each transition rate

Transition rate	Finding	Design Guidance	Example
ALL	• Stationary gaze entropy suggests performance was expected to be worse with more even attention transitions across tasks.	• Conduct a task analysis for the content and placement of AOIs o Minimize transitions between AOIs and/or centralize the most used AOIs o Offload less prioritized tasks to other sensory channels (e.g., chat messages to auditory, fuel leaks to tactile; Riggs et al., 2017)	• Alerting participants of potential threats via the tactile channel improved performance without hindering secondary tasks in a simulated combat environment (Oskarsson, Eriksson, & Carlander, 2012)
SLOW	• Longer fixation duration was predictive of better baseline performance	• Design elements on the display that prompt further examination so the operator is prompted to spend time encoding information • Make items essential to performance, e.g., the target, engaging and informative so the operator spends time fixating (Jacob & Karn, 2003; Poole & Ball, 2006)	• Increasing the information detail and salience without increasing clutter helped refocus attention without a cost of cognitive load in nuclear power plant control environments (Kovesdi et al., 2018)
MEDIUM	• Stationary gaze entropy is the only metric able to predict response time and accuracy trends, so it should be further explored to be used in real-time	• Build and test stochastic models of transitions between AOIs (i.e., Markov chains) to identify the most to least frequent transitions between AOIs to determine the distance between AOIs	• Gaze location in static image viewing was predicted with 56% accuracy (chance was 33%) with hidden Markov modeling (Coutrot, Hsiao, & Chan, 2018)
FAST	• Larger gaze transition rate was predictive of better baseline performance and its trend over time	• Prompt efficient scanning by decluttering the display and making key tasks salient and informative (Moacdieh & Sarter, 2015) • Provide redundancy for the most important tasks by including multiple, informative visual representations in the environment (Yang et al., 2018)	• Comprehension rates and visual attention transition rates increased when information was presented across multiple visualizations (O'Keefe et al., 2014)

Future work and limitations

Although the present work substantially adds to the workload transition knowledgebase, it is not without limitations. First, although the selection of eye tracking metrics included in this study was motivated by Chapter 3, there are other metrics worth exploring. Several scan-based metrics had significant bivariate relations with each unconditional growth curve model, so their usefulness may change based on research goals (e.g., the applicability of spatial destiny and fixation duration in the present work versus Chapter 3). Also, eye tracking metrics used to study cognitive load, such as pupillometry and blink rate, may be better equipped to address any applicability of the resource depletion explanation and/or the recruitment of mental resources due to their ability to address amounts of mental resources, specifically.

Relatedly, multivariate growth curve modeling could explore if the way that scan-based metrics trend over time relates to workload transition performance. It would require a rather large sample size (>100 participants; Astivia et al., 2019), but it may be particularly informative to build upon the resource depletion explanation and/or the recruitment of resources, as both depend on time and workload. Additionally, this approach could include scan-based measures like coefficient *K*, which identifies ambient versus focal visual attention as fixations and saccades occur sequentially over time (Krejtz et al., 2016), and has previously been informative of workload transition performance (Devlin et al., 2021). The current eye tracking metrics were added as time-invariant predictors to the model because of how these metrics are calculated, i.e., measures that were only measured once per person, and the ultimate goal to understand the predictive capability of eye tracking on workload transition performance trends.

Future work should see if Markov models could reliably predict visual attention transitions across AOIs during workload transitions. Previous work has shown some success of using Markov models to predict visual attention allocation (Ebeid et al., 2019; Liechty, Pieters, & Wedel, 2003), but it is unclear how transition rate, individual differences, and task features would impact findings, which is essential in ever-increasing complex environments (Cummings, 2014).

Conclusion

In summary, the present results suggest scan-based metrics can predict workload transition performance trends. They are also very informative on differences in performance trends across transitions rates, as they are able to concretely detail theory and provide design guidance. Future work should continue to explore how novel measures and methods can test and revise theory surrounding workload transitions in order to innovate the current state of the knowledgebase. This work finds another application where studying visual attention allocation is informative of performance trends and shows potential to be applied to operational environments. Future work should build upon the present momentum and include eye tracking and longitudinal data analysis methods to their workload transition research questions as several, relevant follow-up questions are apparent from the present work.

References

Abich, J., Reinerman-Jones, L., & Matthews, G. (2017). Impact of three task demand factors on simulated unmanned system intelligence, surveillance, and reconnaissance operations. *Ergonomics*, *60*(6), 791–809. https://doi.org/10.1080/00140139.2016.1216171

Astivia, O. L. O., Gadermann, A., & Guhn, M. (2019). The relationship between statistical power and predictor distribution in multilevel logistic regression: A simulation-based approach. *BMC Medical Research Methodology*, *19*(1). https://doi.org/10.1186/s12874-019-0742-8

Ayasse, N. D., & Wingfield, A. (2020). The two sides of linguistic context: Eye-tracking as a measure of semantic competition in spoken word recognition among younger and older adults. *Frontiers in Human Neuroscience*, *14*. https://doi.org/10.3389/fnhum.2020.00132

Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474. https://doi.org/10.1016/j.jml.2007.09.002

Barz, M., Kapp, S., Kuhn, J., & Sonntag, D. (2021). Automatic recognition and augmentation of attended objects in real-time using eye tracking and a head-mounted display. *ACM Symposium on Eye Tracking Research & Applications*, 1–4. https://doi.org/10.1145/3450341.3458766

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bowers, M. A., Christensen, J. C., & Eggemeier, F. T. (2014). The effects of workload transitions in a multitasking environment. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 220–224. https://doi.org/10.1177/1541931214581046

Boyer, M., Cummings, M. L., Spence, L. B., & Solovey, E. T. (2015). Investigating mental workload changes in a long duration supervisory control task. *Interacting with Computers*, *27*(5), 512–520. https://doi.org/10.1093/iwc/iwv012

Cerruti, L. M., Estepp, J. R., Miller, W. D., & Christensen, J. C. (2010). Transcranial doppler assessment of workload transition in a complex task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 244–248. https://doi.org/10.1518/107118110X12829369200837

Coutrot, A., Hsiao, J. H., & Chan, A. B. (2018). Scanpath modeling and classification with hidden Markov models. *Behavior Research Methods*, *50*(1), 362–379. https://doi.org/10.3758/s13428-017-0876-8

Cummings, M. (2014). Man versus Machine or Man + Machine? *IEEE Intelligent Systems*, *29*(5), 62–69. www.computer.org/intelligent

Department of Defense. (2017). *Unmanned systems integrated roadmap FY 2017-2042*.

Devlin, S. P., Byham, J. K., & Riggs, S. L. (2021). Does what we see shape history? Examining workload history as a function of performance and ambient/focal visual attention. *Transactions on Applied Perception*, *8*(2), 1-17. https://doi.org/10.1145/3449066

Ebeid, I. A., & Gwizdka, J. (2018). Real-time gaze transition entropy. *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 1-3. https://doi.org/10.1145/3204493.3208340

Feitshans, G., Rowe, A., Davis, J., Holland, M., & Berger, L. (2008). Vigilant spirit control station (VSCS) the face of counter. *AIAA Guidance, Navigation and Control Conference and*

Exhibit. https://doi.org/10.2514/6.2008-6309

Foroughi, C. K., Sibley, C., Brown, N. L., Rovira, E., Pak, R., & Coyne, J. T. (2019). Detecting automation failures in a simulated supervisory control environment. *Ergonomics*, *62*(9), 1150-1161. https://doi.org/10.1080/00140139.2019.1629639

Gazepoint Control user manual. (2019). https://www.gazept.com/

Gluckman, J. P., Warm, J. S., Dember, W. N., & Rosa, R. R. (1993). Demand transitions and sustained attention. *The Journal of General Psychology*, *120*(3), 323–337. https://doi.org/10.1080/00221309.1993.9711151

Godfroid, A., Ahn, J., Choi, I., Ballard, L., Cui, Y., Johnston, S., Lee, S., Sarkar, A., & Yoon, H. J. (2018). Incidental vocabulary learning in a natural reading context: An eye-tracking study. *Bilingualism*, *21*(3), 563–584. https://doi.org/10.1017/S1366728917000219

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: Methods and constructs. *International Journal of Industrial Ergonomics*, *24*(1), 631–645. https://doi.org/10.1016/S0169-8141(98)00068-7

Gudivada, V. N., Rao, D., & Raghavan, V. V. (2015). Chapter 9: Big data driven natural language processing research and applications. In V. Govindaraju, V. V. Raghavan, & C. R. Rao (Eds.), *Handbook of Statistics* (pp. 203–238). Elsevier. https://www.sciencedirect.com/science/article/pii/B9780444634924000095

Hobbs, A., & Shively, R. J. (2014). Human factor challenges of remotely piloted aircraft. *31st EAAP Conference*, 5–14.

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive energetical framework. *Biological Psychology*, *45*, 73–93. https://doi.org/10.1016/S0301-0511(96)05223-4

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. Routledge.

Hoffman, L. (2021). *Time-invariant predictors in longitudinal models*. https://www.lesahoffman.com/PSQF7375_Longitudinal/PSQF7375_Longitudinal_Lecture7b_Time-Invariant.pdf

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods, paradigms, and measures* (2nd ed). CreateSpace Independent Publishing Platform.

Jacob, R. J. K., & Karn, K. S. (2003). Eye tracking in human–computer interaction and usability research: Ready to deliver the promises. In *In the Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research* (pp. 573–607).

Kim, N., House, R., Yun, M. H., & Nam, C. S. (2019). Neural correlates of workload transition in Multitasking: An ACT-R model of hysteresis effect. *Frontiers in Human Neuroscience, 12*(535), 1–14. https://doi.org/10.3389/fnhum.2018.00535

Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Koh, D. H., & Gowda, S. M. (2010). Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering*, *57*(11), 2635–2645. https://doi.org/10.1109/TBME.2010.2057429

Komogortsev, O. V., & Karpov, A. (2013). Automated classification and scoring of smooth pursuit eye movements in the presence of fixations and saccades. *Behavior Research Methods*, *45*(1), 203–215. https://doi.org/10.3758/s13428-012-0234-9

Kovesdi, C., Spielman, Z., Leblanc, K., & Rice, B. (2018). Application of eye tracking for measurement and evaluation in human factors studies in control room modernization.

Nuclear Technology, *202*(2–3), 220–229. https://doi.org/10.1080/00295450.2018.1455461

Krejtz, K., Duchowski, A., Szmidt, T., Krejtz, I., González Perilli, F., Pires, A., Vilaro, A., & Villalobos, N. (2015). Gaze transition entropy. *ACM Transactions on Applied Perception*, *13*(1), 1–20. https://doi.org/10.1145/2834121

Krejtz, K., Duchowski, A., Krejtz, I., Szarkowska, A., & Kopacz, A. (2016). Discerning ambient/focal attention with coefficient *K*. *ACM Transactions on Applied Perception*, *13*(3), 1–20. https://doi.org/10.1145/2896452

Liechty, J., Pieters, R., & Wedel, M. (2003). Global and local covert visual attention: Evidence from a bayesian hidden Markov model. *Psychometrika*, *68*(4), 519–541. https://doi.org/10.1007/BF02295608

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology*, *97*(5), 951–966. https://doi.org/10.1037/a0028380

McKendrick, R., & Harwood, A. (2019). Cognitive workload and workload transitions elicit curvilinear hemodynamics during spatial working memory. *Frontiers in Human Neuroscience*, *13*(405), 1–16. https://doi.org/10.3389/fnhum.2019.00405

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494. https://doi.org/10.1016/j.jml.2007.11.006

Moacdieh, N. M., Devlin, S. P., Jundi, H., & Riggs, S. L. (2020). Effects of workload and workload transitions on attention allocation in a dual-task environment: Evidence from eye tracking metrics. *Journal of Cognitive Engineering and Decision Making*, *14*(2), 132–151. https://doi.org/10.1177/1555343419892184

Moacdieh, N., & Sarter, N. (2015). Display clutter: A review of definitions and measurement techniques. *Human Factors*, *57*(1), 61–100. https://doi.org/10.1177/0018720814541145

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, *42*(1), 188–204. https://doi.org/10.3758/BRM.42.1.188

O'Keefe, P. A., Letourneau, S. M., Homer, B. D., Schwartz, R. N., & Plass, J. L. (2014). Learning from multiple representations: An examination of fixation patterns in a science simulation. *Computers in Human Behavior*, *35*, 234–242. https://doi.org/10.1016/j.chb.2014.02.040

Oskarsson, P. A., Eriksson, L., & Carlander, O. (2012). Enhanced perception and performance by multimodal threat cueing in simulated combat vehicle. *Human Factors*, *54*(1), 122–137. https://doi.org/10.1177/0018720811424895

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., Heisterkamp, S., Willigen, V. B., & Ranke, J. (2021). *Package "nlme."* https://bugs.r-project.org

Poole, A., & Ball, L. J. (2006). Eye tracking in human-computer interaction and usability research: Current status and future prospects. *Encyclopedia of Human-Computer Interaction*, 211–219. https://doi.org/10.4018/978-1-59140-562-7.ch034

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. SAGE.

Ratwani, R. M., McCurry, J. M., & Trafton, J. G. (2010). Single operator, multiple robots: An eye movement based theoretic model of operator situation awareness. *HRI '10 Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, *55*, 235–242. https://doi.org/10.1109/HRI.2010.5453191

Rizopoulos, D. (2021). Package 'GLMMadaptive''.' In *Journal of Computational and Graphical Statistics*, *41*(1), https://doi.org/10.1080/10618600.1995.10474663

Riggs, S. L., Wickens, C. D., Sarter, N., Thomas, L. C., Nikolic, M. I., & Sebok, A. (2017). Multimodal information presentation in support of NextGen operations. *International Journal of Aerospace Psychology*, *27*(1–2), 29–43. https://doi.org/10.1080/10508414.2017.1365608

RStudio Team (2020). *RStudio: Integrated Development for R.* RStudio, PBC, Boston, MA URL http://www.rstudio.com/

Shic, F., Chawarska, K., Bradshaw, J., & Scassellati, B. (2008). Autism, eye-tracking, entropy. *7th IEEE International Conference on Development and Learning, ICDL*, 73–78. https://doi.org/10.1109/DEVLRN.2008.4640808

Shiferaw, B., Crewther, D., & Downey, L. A. (2019). Gaze entropy measures reveal alcohol-induced visual scanning impairment during ascending and descending phases of intoxication. *Journal of Studies on Alcohol and Drugs*, *80*(2), 236–244. https://doi.org/10.15288/jsad.2019.80.236

Shiferaw, B., Downey, L. A., Westlake, J., Stevens, B., Rajaratnam, S. M. W., Berlowitz, D. J., Swann, P., & Howard, M. E. (2018). Stationary gaze entropy predicts lane departure events in sleep-deprived drivers. *Scientific Reports*, *8*(1), 1–10. https://doi.org/10.1038/s41598-018-20588-7

Sibley, C., Coyne, J., & Morrison, J. (2015). Research considerations for managing future unmanned systems. *AAAI Spring Symposium*, 54–59.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Steichen, B., Carenini, G., & Conati, C. (2013). User-adaptive information visualization - using eye gaze data to infer visualization tasks and user cognitive abilities. *IUI '13*, 19–22.

Wickens, C. D. (1980). The structure of attentional resources. In R. S. Nickerson (Ed.), *Attention and Performance VIII* (pp. 239–257). Psychology Press.

Yang, F.-Y., Tsai, M.-J., Chiou, G.-L., Wen-Yu Lee, S., Chang, C.-C., & Chen, L.-L. (2018). Instructional suggestions supporting science learning in digital environments based on a review of eye tracking studies. *Source: Journal of Educational Technology & Society*, *21*(2), 28–45. https://doi.org/10.2307/26388377

CHAPTER 6

Conclusion

Workload transitions continue to be an open, relevant research topic, for settings like UAV command and control given the features of the environment are constantly changing and the expectations of the operator are expanding (Cummings et al., 2014, 2019; Sibley et al., 2015). This dissertation addresses some of the research gaps by concurrently considering performance measures with eye tracking data. Specifically, this dissertation addresses:

- The impact transition rate, i.e., the speed at which workload changes from one load to another, has on multitasking performance and performance over time.
- The ability of eye tracking data to explain and predict the observed performance trends of workload transitions.
- The empirical development of the current theoretical explanations as a means to inform future display design for the dynamic environments.

The present research finds workload transition performance depends on the task, transition rate, and individual. However, visual attention allocation patterns are valuable in explaining and predicting these effects. Specifically, Chapters 2 and 3 shed light on how people multitask and adjust visual attention allocation patterns as workload transitions. Chapters 4 and 5 confirm and detail how transition rate influences performance and visual attention allocation and offers amendments to an existing theoretical explanation of workload transitions. Finally, this work proposes display design guidance to account for transition rate, so that technology supports operators in environments prone to workload transitions, like UAV command and control.

Intellectual Merit: Contributions to the Workload Transition Knowledgebase

This dissertation addresses a specific, unexplored niche in the knowledgebase: the study of how visual attention allocation evolves under different types of workload transitions in multitasking environments. Specifically, the present dissertation builds upon the following aspects of workload transitions:

- Existing theoretical explanations,
- Human performance modeling, and
- The predictive and explanatory value of visual attention allocation patterns.

It also lays the groundwork for human factors professionals to pursue different analysis techniques for their multifaceted, operationally relevant research questions.

Building upon current workload transition theory: The effort regulation explanation

The workload transition knowledgebase is fragmented, with no single theory unifying the observed performance trends. The most consequential finding of this dissertation is the contribution it makes towards the theoretical explanations surrounding workload transitions. Specifically, the effort regulation explanation, which states performance during workload transitions is dependent on the ability to appraise, recruit, and deploy mental resources effectively (Hockey, 1997). The following findings from each chapter support and add to the effort regulation explanation:

- **Chapter 2**: Performance was better during workload transitions than constant low workload and remained stable or recovered with multiple occurrences of workload transitions.

- **Chapter 3**: Better multitasking performance was associated with more concentrated and direct visual attention and primary task performance was better when visual attention type followed workload level expectations over time, i.e., focal during high workload and ambient during low workload.
- **Chapter 4**: Primary task performance trends were different for each transition rate, and even within a given transition rate, as performance either improved, stagnated, or worsened over time.
- **Chapter 5**: Scan-based eye tracking metrics predicted primary task performance trends over time for each transition rate. Specifically, attention transitions that were more dispersed predicted worst primary task performance.

Chapter 2 supports the effort regulation explanation because improved performance was not only possible during constant low levels of workload, as performance improved with high workload and multiple workload transitions for both the primary and secondary task. Chapter 3 lent to the effort regulation explanation because better performance was associated with *how* visual attention was used, specifically in terms of spread (*where are users generally looking?*) and directness (*how purposeful are attention transitions?*). The follow up studies (Chs. 4 and 5) provided evidence to amend the effort regulation explanation to say that performance during workload transitions is dependent on the ability to appraise both the workload level *and* transition rate, as well as deploy *visual* attention resources effectively, particularly in terms of spread, directness, and duration. These studies highlight the nuances associated with adaptation-based models of mental resources (Hancock & Warm, 1989; Hockey, 1986), but providing specifications helps understand their applicability. For example, the Multiple Resource Model, which posits that different sensory modalities draw from separate attentional resources, specifies how multitasking performance is a function of relying on mental resource from different sensory channels and information stages (Wickens, 1980). Providing these specifications has led to a more informed understanding and expectations of multitasking performance across applications (Wickens, 2008). There are still unanswered details surrounding the effort regulation explanation, but this work begins to shed light on the how mental resources are used during workload transitions and the factors that influence their usage. This is of particular importance when trying to clarify the nuance surrounding workload transition performance.

The consistent support for the effort regulation explanation actually gives a rationale for the performance differences observed in the workload transition literature. The results from Chapters 4 and 5 suggest differences stem from an *individual's* visual attention allocation patterns. This not only builds on the notion of individual differences being influential to workload transition performance (Cox-Fuenzalida et al., 2004, 2006; McKendrick & Harwood, 2019; Mracek et al., 2014), but it provides an explanation as to *why* – i.e., individuals' visual attention allocation patterns. Chapter 4 shows the value of considering the individual's performance trends, so future work should continue to do so; otherwise there should be a cautious interpretation of the results as basing decisions only on pairwise comparisons may be problematic and/or incomplete. Given the importance of visual attention allocation patterns have on workload transition performance (Chs. 3 and 5), measures of an individual's attentional control should be considered in future workload transition research (Engle, 2018). Also, technology and systems should be more customized to the individual (Szalma, 2009) in the hopes of making them more effective tools for each operator managing dynamic environments.

Studying workload transitions with context adds to human performance modeling

Human performance models for workload transitions should be expanded to directly account for *how* transitions occur within the context of the environment, e.g., the transition rate (Chs. 2 and 4), and how it may not impact all individuals in the same way, e.g., their performance trends over time (Ch. 4) or their visual attention allocation patterns (Ch. 5).

- **Chapter 2**: Workload history effects of medium transitions were more severe compared to the workload history effects of fast transitions. However, secondary task performance was the same for the two transition rates.
- **Chapter 4**: Primary task performance trends over time depend on the transition rate, performance measure, workload level, and the individual. Slow transitions lead to more stable performance across each workload period, while medium transitions show the potential to improve performance for some individuals over time. Fast transitions lead to some of the fastest performance, with a minimal cost to accuracy.
- **Chapter 5**: Workload transition rate influences visual attention allocation patterns differently.

Given this dissertation is one of the first thorough investigations of workload transition rate, the workload transition literature is inherently expanded from the present results. It also was essential to specifying theory, which not only increases the value of its impact, but also suggests it needs to be considered in future research.

Transition rate should ideally be accounted for in human performance models. For instance, it should be added as a component of S-PRINT, a human performance model that predicts how performance will fare when workload suddenly increases to levels of overload after extended periods of underload (Sebok et al., 2017). There were some instances where the performance differences amongst transition rates would lead to very different practical outcomes. For example, Chapter 2 finds primary task accuracy worsens and never recovers during medium transitions, but remains stable during fast transitions. Performance also could differ within a given transition rate, as Chapter 4 finds medium transitions lead to improved performance for some participants and not others. Chapter 3 showed that workload transitions affected how visual attention was allocated and Chapter 5 showed how it depended on the transition rate. Based on these collective findings, we cannot assume workload transitions impact operators in the same way. Instead, transition rate and visual attention allocation, needs to be actively accounted for when modeling workload transition performance. Currently, S-PRINT does not include features of the workload transition as a model component, and rather focuses on more general operational aspects, like operator's fatigue level and human-automation interaction. S-PRINT could benefit from adding specifics related to workload transitions in order to increase its accuracy and utility.

More broadly, human performance models of multitasking should consider how workload transitions may impact task completion strategies by better understanding how tasks priorities fluctuate in the environment (Clare et al., 2010). Chapter 3 found workload transitions made visual attention allocation more efficient and direct and this benefitted both primary and secondary task performance (Ch. 2). Currently, the Strategic Task Overload Management model (STOM; Wickens et al., 2013) is based on the idea that task completion is largely dictated by its engagement, as defined by its urgency, saliency, difficulty. This dissertation suggests the dynamics of the environment, i.e., its propensity to shift workload and the interdependency of the tasks, should also be considered given its rather profound impact on multitasking performance trends. The results of this dissertation also directly emphasize the recent calls to explore how the time on task (Ch. 4), fluctuations in task priority (Ch. 2), and individual differences (Ch. 4) impact task completion and therefore STOM (Wickens & Gutzwiller, 2017), as all were influential on workload transition performance.

Overall, the current results support the general need for human performance models as features of the environment, task, and their interaction appear to greatly influence task completion strategies in dynamic environments. Focusing on relevant features (e.g., transition rate) with as much context as possible (e.g., dynamic multitasking, sampling from the relevant population, etc.) was successful in contributing to the knowledgebase, which supports the growing call for human factors research to be conducted with increased specificity and context for greater impact (Dul et al., 2012). Clearly, performance trends are subject to *how* certain events occur the environment and *how* they are managed by the individual. With the growing complexity of environments (Cummings, 2014), it is essential to include as many realistic features as possible in research initiatives and continuously iterate in identifying and studying influential features.

Visual attention allocation with workload transitions

This dissertation also makes a substantial impact on a specific, unexplored niche in the knowledgebase: the study of how visual attention allocation evolves over time under workload transitions in a multitasking environment. Specifically, it finds:

- **Chapter 3**: Compared to constant workload, workload transitions leads to visual attention transition to be more concentrated and efficient, and visual attention allocation patterns develop over time.

- **Chapter 5**: Across all transitions rates, better performance was (again) associated with less dispersed visual attention transitions. However, there were still notable differences in visual attention allocation patterns for each transition rate, which lend to explain and predict their currently observed performance differences over time.

This research is among the few studies to show that psychophysiological measures can predict certain aspects of workload transition performance via growth curve modeling (Kim et al., 2019). Visual attention allocation newly identified *why* individuals have varying levels of success during a workload transition and what that means for display design (Ch. 3 and 5). Questions addressed by studying visual attention allocation include: why multitasking performance was better than constant workload (Ch. 3), why performance recovered for a certain transition rate (Ch. 3), or why performance varied across individuals (Ch. 5). More generally, the visual attention allocation patterns suggested performance during workload transitions may not depend on the amount of mental resources, but rather how those resources are organized and used, which builds upon a suite of theories suggesting performance is not as dependent on the sheer amount of workload, but rather how the task prompts mental resource usage (Abich et al., 2017; Hockey, 1997; Sebok et al., 2015; Wickens, 2002; Young & Stanton, 2002).

This work also confirms the value of advanced scan-based measures—i.e., gaze transition entropy, stationary gaze entropy, and coefficient *K*—as they were essential in explaining and/or predicting performance outcomes. This dissertation supports the notion that higher-level descriptors of visual attention allocation are needed to understand cognitive processes in dynamic environments (Krejtz et al., 2016). Advanced metrics can provide the nuance and specificity other methods struggle to provide, (e.g., think aloud protocol, debriefing strategies, etc.), further confirming the benefit of continuous, unobtrusive, psychophysiological measures. For example, Abich et al. (2017) note the need to quantify and compare how dispersed attention was across the display, but how it would be extremely challenging with debriefing data alone. Continuously monitoring and understanding the management process used by the individual in these environments is not only useful for prediction (Ch. 5), but it can provide very tangible and specific insights on convoluted concepts, such as the nuance and variation in workload transition performance, e.g., the slight differences observed in coefficient *K* over time in Chapter 3.

New analysis approaches to be considered by human factors researchers and practitioners

When trying to provide clarity for an underdeveloped or poorly understood human factors concept, multiple investigations, measures, and analyses may be necessary. This dissertation not only shows the value, but also a method, on how to thoroughly study inherent features of an environment and synthesize across different measures. For example:

- **Chapter 2:** Expanding both the performance and eye tracking analysis to compare across workload periods was more meaningful when studying transition rate.
- **Chapter 3:** Advanced scan-based metrics lead to insightful findings towards understanding human performance over time.
- **Chapter 4:** Conducting both the aggregate analysis and growth curve modeling for response time and accuracy of all transition rates lead to operator-centric design recommendations.
- **Chapter 5:** Directly applying scan-based metrics as predictors of individual's performance led to some of the most groundbreaking findings of this dissertation.

Developing multifaceted experimental designs and thorough analysis approaches are only increasing in importance as these environments continue to increase in complexity and potentially threaten operator and system safety (Cummings, 2014). Making minor adjustments to central parts of the research question in follow-up investigations (e.g., expanding the transition rate investigation, Ch. 4), relying on different measures (e.g., eye tracking Chs. 3 and 5), and analysis approaches (e.g., growth curve modeling, Chs. 4 and 5) not only confirmed the magnitude of the observed effect, but it also led to more nuanced prediction models, theory development, and design guidance.

Human factors work should ideally consider using multiple analysis approaches when exploring relevant topics to dynamic domains. This notion is directly supported when reviewing the outcomes of the aggregate and growth curve modeling results in Chapter 4. The interpretation from each analysis differed as the aggregate analysis suggested slow transitions lead to some of the most stable performance, but growth curve modeling estimated performance to improve during medium transitions. If conclusions were only based on the typical aggregate analysis, then theory and design guidance may have only been applicable for “average” performers – a very small portion of the population that becomes even more irrelevant as UAVs span to different military populations (Freedberg, 2018). Topics like situation awareness and fatigue should explore how the synthesis of longitudinal models and psychophysiological measures can address their current research gaps, given these topics can evolve over time, be physiologically tracked, and vary across individuals in dynamic environments (Endsley, 2015; Guastello et al., 2012; Salfinger et al., 2013).

Broader Impacts: Informing the Design of Technology in UAV Environments

This research occurred in the context of the complex and dynamic domain of UAV command and control. However, the experimental design (e.g., studying transition rate, including psychophysiological measures of the operator, sampling from a context-relevant population, etc.) and analysis approach (e.g., longitudinal models of workload transition performance with psychophysiological data as model predictors), can benefit various complex, dynamic work environments (e.g., aviation, nuclear plant control, and emergency response; Huey & Wickens, 1993). This dissertation delivers empirically based guidelines on display design as a means to better account for an operator's needs in real-time during workload transitions. It also expands the methodological options for adaptive displays (i.e., displays that adjust the information presentation, content, or amount based on the real-time needs of the operator; Feigh et al., 2012).

Informing display design in complex work environments

An immediate applied takeaway from this work is the understanding of how to design displays to account for workload transitions. Specifically, each chapter finds:

- **Chapter 2:** Workload transitions may change strategies of multitasking.
- **Chapter 3:** To improve multitasking performance, display should be designed so that visual attention transitions are efficient and concentrated to a few necessary areas on the display.
- **Chapter 4:** Depending on the performance goals, technology and systems should be designed to foster certain types of workload transitions.

- **Chapter 5:** Tasks and information essential to performance should be collocated on the display.

Across all transition rates, performance improved when individuals dedicated their visual attention less equally across the display (Chs. 3 and 5). Performance improved further if individuals adjusted their visual attention strategy for the transition rate accordingly (Ch. 5). Although these scan patterns may be trainable (Vine et al., 2012), it may be more advantageous to design displays that prompt visual attention allocation accordingly. For example, features of the display that are essential to system performance or safety should be collocated. Display design should also be based on the transition rate (Table 5.8 in Ch. 5), which can vary per the performance goals of the environment (Table 4.8 in Ch. 4). UAV command and control environments need extensive research on display design because their operators sometimes primarily rely on, or only have, visual information during missions (Hobbs & Shivley, 2014). They also rely on automation with unprecedented functionality (e.g., multiple, related autonomous systems; Cummings et al., 2019). This creates a very different environment when compared to manned aerial missions (McCarley & Wickens, 2004), so simply applying that display guidance may not suffice. Display design may not only prevent or mitigate large-impact errors, it might also reduce the burnout amongst operators working incredibly long shifts in these high-stakes missions (Arrabito et al., 2010). Burnout can lead to serious personal health issues and long-term staffing problems, which has been identified as a threat to the UAV population specifically (Ouma et al., 2011), but is by no means unique to this population (e.g., healthcare; Moss et al., 2016; van Wulfften Palthe et al., 2016). Preventing and tempering burnout is not new to human factors research either, so further investigations of this kind are warranted (Matthews et al., 2019). Informing display design with a multifaceted approach, i.e., simultaneously studying the impact specific features of the task, environment, and individual have on performance, also informs aspects of the system as well, including the open question of selection and training procedures (Harkins, 2020).

Laying the foundation for adaptive displays: Using scan-based metrics and novel analysis methods

Given the goal is to implicitly monitor the real-time state of the operator without imposing any additional burden on that operator (Feigh et al., 2012), it is important to identify adaptive drivers, i.e., measures that are able to continuously and implicitly monitor the operator's state and predict their needs in real-time (Rothrock et al., 2002). This dissertation finds:

- **Chapter 3:** Measuring how dispersed and direct visual attention was, especially in terms of how it related to the tasks, helped identify differences in multitasking performance between transitioning and constant workload.
- **Chapter 5:** Understanding how efficiency and duration of visual attention explained the performance differences between individuals and transition rate.

Scan-based metrics show promise to be adaptive drivers because they predicted workload transition performance across different transition rates and performance measures (Ch. 5). Even more promising, they suggested *when* and *how* visual attention patterns should adjust for the transition rate and performance goals, which is key to effective adaptive assistance. The current results suggest adaptive drivers should monitor the distribution of attention switches across the display and prompt the operator to reduce the frequency of those switches when target detection rates need to improve (Chs. 3 and 5). Growth curve modeling should be further explored as the basis of adaptive displays as it not only identified the differences in individual performance, but also how visual attention allocation patterns contributed to those differences across difference transition rates (Ch. 5).

Given the current success of gaze transition entropy and stationary gaze entropy (Chs. 3 and 5), another potential alternative for adaptive display algorithms is basing them on Markov models. It is becoming increasingly necessary to explore and verify stochastic models for human performance, as the true characteristics of human behavior continue to be observed and characterized as naturally dynamic (Feng et al., 2016; Hancock & Matthews, 2019). This dissertation shows the promise of modeling visual attention patterns as a Markovian process, which further supports some our preliminary investigations (Devlin & Riggs, 2017). Practically, it may be successful in predicting the location of visual attention in real-time in a dynamic environment, which is an essential first step in catering to the operator's most pressing needs. This alternative also has a practical benefit, as it would require minimal input and training data, which is a drawback of methods like growth curve modeling and machine learning (Delucia & Pitts, 2006; Kruthiventi, Ayush, & Babu, 2017). This benefit is particularly relevant for eye tracking considering the inevitable loss of eye tracking data in real-time (Holmqvist et al., 2011; Sibley et al., 2017). Providing adaptive assistance has innovated several fields, most notably in the learning and training domain (e.g., Bayesian Knowledge Tracing; Corbett & Anderson, 1994). Ever since the implementation of intelligent tutoring systems, students have learned faster and retained the information longer, which inherently benefits society (Chassignol et al., 2018). Similar societal benefits may follow if we move away from post hoc methods of design (Hobbs & Shivley, 2014) and towards proactively designing displays to thwart and predict operator error in UAV command and control, as a means to substantially and sustainably reduce consequential mishaps and accidents (Breslow et al., 2014; Williams, 2006).

Future Work

This dissertation simultaneously expands and details the workload transition knowledgebase, specifically by better informing theoretical explanations and the design of displays in complex and dynamic environments. It also sets the stage for several future research efforts, ranging from direct follow-up investigations to new expansion on promising findings. The benefits of completing this future work include expanding theory, human performance models, and technology design for workload transitions.

A follow-up study to the current work would be to explore how the number and/or duration of workload transitions affects performance in a similar, dynamic, multitasking setting. It is currently unclear if the instances where performance stabilized was due to transition rate and/or expectancy effects (i.e., awareness of/preparing for workload to shift given it had several times prior; Kochan et al., 2004; Landman et al., 2017). The only other previous work that studies multiple workload transitions also finds performance stabilized as workload transitions multiple times in the experiment, but they only ever transitioned workload instantly (Morgan & Hancock, 2011). Completing this work could address the validity of the resource depletion explanation. If performance trends are similar even when the number of transitions differs, this would not support the resource depletion explanation as resources recover only when workload is low. Further exploring the impact of multiple workload transitions would inform if technology needs to intervene for every instance of a workload transition or if there is any benefit to have the operator experience workload transitions under certain circumstances.

There is also a need to understand how different, occupationally-relevant modulations of workload transitions, especially ones that span sensory modality, influence performance. Relying on scan-based metrics for an entirely visual task was greatly successful in explaining nuanced performance trends surrounding workload transitions. Future work should continue to include real-time cognitive measures for both theoretical and practical applications. For example, it may be worthwhile to include additional psychophysiological measures as task demands span sensory modality. For example, heart rate measures should be included if tasks rely on the auditory channel (see review in Erfanian et al., 2019). Although using multiple psychophysiological measures within the same experiment can fail to converge to the same conclusions (Matthews et al., 2015), initial investigations are necessary to better understand the operator's experience of multimodal workload transitions.

Future work should also actively search and identify the specific features of the individual that are moderating workload transition performance. This dissertation calls for future research to go beyond identifying and/or controlling for individual differences, as it should scrutinize how features of the individual influence performance. For example, eye tracking studies find training novices with expert scan patterns improves performance and learning rates (Law et al., 2005; Vine et al., 2012), so conducting this type of investigation will directly inform what to expect from and how to design for the individual in environments where workload transitions differently. Along the same lines, understanding how teams of individuals perform when workload transitions in these environments will be critical across complex domains. Developing measures and methods that specifically monitor and assist with effective collaboration (e.g., understanding the effect of team personality, task completion strategy, and quantifying their shared visual attention allocation patterns with new eye tracking metrics and analysis methods; e.g., Devlin et al., 2018; Devlin et al., 2019; Devlin et al., 2020a, 2020b) is essential to further understanding workload transitions and display design alike.

To confirm and expand upon the value of eye tracking data, future work should consider other eye tracking metrics. Here, several scan-based metrics predicted performance trends when it was the only predictor in the growth curve model and the applicability of these measures can change per research goals (e.g., the applicability of spatial destiny and fixation duration in Ch. 3 vs. 5). It is unrealistic to assume a single set of scan-based metrics will be applicable to all research situations, but this work can serve as a starting point for future explorations focused on scan-based metrics. Also, it may be interesting to include eye tracking metrics for cognitive load (e.g., pupillometry, blink rate). These measures may be better equipped to address any applicability of the resource depletion explanation and/or the recruitment process of mental resources due to their specific capability to measure amounts of mental resources.

Finally, different analysis techniques should also be considered. Multivariate growth curve modeling could explore if eye tracking metrics evolve over time and if/how this relates to workload transition performance. This would require a larger sample size (i.e., >100; Astivia et al., 2019), but it may be particularly informative to build upon the resource depletion explanation and/or the recruitment of resources, as both depends on time and workload. Additionally, multivariate growth curve modeling could allow for the inclusion of coefficient *K*, given its dependence on timing of fixations and saccades and present workload level (Ch. 3). Eye tracking metrics were added as time-invariant predictors in this dissertation because of how these metrics are calculated, i.e., measures that were only measured once per person, and the ultimate goal to understand the predictive capability of eye tracking on performance trends.

References

Abich, J., Reinerman-Jones, L., & Matthews, G. (2017). Impact of three task demand factors on simulated unmanned system intelligence, surveillance, and reconnaissance operations. *Ergonomics*, *60*(6), 791–809. https://doi.org/10.1080/00140139.2016.1216171

Astivia, O. L. O., Gadermann, A., & Guhn, M. (2019). The relationship between statistical power and predictor distribution in multilevel logistic regression: A simulation-based approach. *BMC Medical Research Methodology*, *19*(1). https://doi.org/10.1186/s12874-019-0742-8

Clare, A. S., Hart, C. S., & Cummings, M. L. (2010). Assessing operator workload and performance in expeditionary multiple unmanned vehicle control. *48th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition*, 1–10. https://doi.org/10.2514/6.2010-763

Cox-Fuenzalida, L.-E., Angie, A., Holloway, S., & Sohl, L. (2006). Extraversion and task performance: A fresh look through the workload history lens. *Journal of Research in Personality*, *40*(4), 432–439. https://doi.org/10.1016/j.jrp.2005.02.003

Cox-Fuenzalida, L.-E., Swickert, R., & Hittner, J. B. (2004). Effects of neuroticism and workload history on performance. *Personality and Individual Differences*, *36*(2), 447–456. https://doi.org/10.1016/S0191-8869(03)00108-9

Cummings, M. (2014). Man versus machine or man + machine? *IEEE Intelligent Systems*, *29*(5), 62–69. www.computer.org/intelligent

Cummings, M. L., Huang, L., Zhu, H., Finkelstein, D., & Wei, R. (2019). The impact of increasing autonomy on training requirements in a UAV supervisory control task. *Journal of*

Cognitive Engineering and Decision Making, *13*(4), 295–309. https://doi.org/10.1177/1555343419868917

DeLucia, C., & Pitts, S. C. (2006). Applications of individual growth curve modeling for pediatric psychology research. *Journal of Pediatric Psychology*, *31*(10), 1002–1023. https://doi.org/10.1093/jpepsy/jsj074

Devlin, S. P., Flynn, J. R., & Riggs, S. L. (2018). Connecting the Big Five taxonomies: Understanding how individual traits contribute to team adaptability under workload transitions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 119–123. https://doi.org/10.1177/1541931218621027

Devlin, S. P., Flynn, J. R., & Riggs, S. L. (2019). Examining the visual attention of pairs of operators during a low to high workload change. *Proceedings of the Human Factors and Ergonomics Society 2019 Annual Meeting*, 201–205. https://doi.org/10.1177/1071181319631168

Devlin, S. P., Flynn, J. R., & Riggs, S. L. (2020a). How coordinated visual attention on a target area of interest is impacted by a change in workload over time. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 366–370. https://doi.org/10.1177/1071181320641083

Devlin, S. P., Flynn, J. R., & Riggs, S. L. (2020b). How shared visual attention patterns of pairs unfold over time when workload changes. *Eye Tracking Research and Applications Symposium (ETRA)*, 41, *1-5*. https://doi.org/10.1145/3379156.3391339

Devlin, S. P., & Riggs, S. L. (2017). Analyzing eye tracking data using a Markovian framework to assess differences in scan patterns. *Proceedings of the Human Factors and Ergonomics Society*, 1814–1818. https://doi.org/10.1177/1541931213601935

Dul, J., Bruder, R., Buckle, P., Carayon, P., Falzon, P., Marras, W. S., Wilson, J. R., & van der Doelen, B. (2012). A strategy for human factors/ergonomics: Developing the discipline and profession. *Ergonomics*, *55*(4), 377–395. https://doi.org/10.1080/00140139.2012.661087

Erfanian, M., Mitchell, A. J., Kang, J., & Aletta, F. (2019). The psychophysiological implications of soundscape: A systematic review of empirical literature and a research agenda. *International Journal of Environmental Research and Public Health*, *16*(19). https://doi.org/10.3390/ijerph16193533

Endsley, M. R. (2015). Final reflections: Situation awareness models and measures. *Journal of Cognitive Engineering and Decision Making*, *9*(1), 101–111. https://doi.org/10.1177/1555343415573911

Engle, R. W. (2018). Working memory and executive attention: A revisit. *Association for Psychological Science*, *13*(2), 190–193. https://doi.org/10.1177/1745691617720478

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. https://doi.org/10.1177/0018720812443983

Freedberg Jr., S. J. (2021, June 7). Drone refuels manned aircraft for first time: Navy. *Breaking Defense*. https://breakingdefense.com/2021/06/drone-refuels-manned-aircraft-for-first-time-navy/

Guastello, S. J., Boeh, H., Shumaker, C., & Schimmels, M. (2012). Catastrophe models for cognitive workload and fatigue. *Theoretical Issues in Ergonomics Science*, *13*(5), 586–602. https://doi.org/10.1080/1463922X.2011.552131.

Harkins, G. (2020, December 22). The Navy wants to recruit 450 warrant officers to fly its new MQ-25 tanker drones. *Military.Com*.

Hobbs, A., & Shively, R. J. (2014). Human factor challenges of remotely piloted aircraft. *31st EAAP Conference*, 5–14.

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive energetical framework. *Biological Psychology*, *45*, 73–93. https://doi.org/10.1016/S0301-0511(96)05223-4

Huey, B. M., & Wickens, C. D. (Eds.). (1993). *Workload transition: Implications for individual and team performance.* National Research Council.

Kim, N., House, R., Yun, M. H., & Nam, C. S. (2019). Neural correlates of workload transition in Multitasking: An ACT-R model of hysteresis effect. *Frontiers in Human Neuroscience*, *12*(535), 1–14. https://doi.org/10.3389/fnhum.2018.00535

Kochan, J. A., Breiter, E. G., & Jentsch, F. (2004). Surprise and unexpectedness in flying: Database reviews and analyses. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 335–339. https://doi.org/10.1177%2F154193120404800313

Krejtz, K., Duchowski, A., Krejtz, I., Szarkowska, A., & Kopacz, A. (2016). Discerning ambient/focal attention with coefficient *K*. *ACM Transactions on Applied Perception*, *13*(3), 1–20. https://doi.org/10.1145/2896452

Kruthiventi, S. S. S., Ayush, K., & Babu, R. V. (2017). DeepFix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, *26*(9), 4446–4456. https://doi.org/10.1109/TIP.2017.2710620

Landman, A., Groen, E. L., van Paassen, M. M., René., Bronkhorst, A. W., & Mulder, M. (2017). The influence of surprise on upset recovery performance in airline pilots. *International Journal of Aerospace Psychology*, *27*(1–2), 2–14. https://doi.org/10.1080/10508414.2017.1365610

Law, B., Atkins, S., Kirkpatrick, A. E., & Lomax, A. (2004). Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment*: Proceedings of the 2004 Symposium on Eye Tracking Research & Applications*, *1*(212), 41–48. https://doi.org/10.1145/968363.968370

Matthews, G., De Winter, J., & Hancock, P. A. (2019). What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. *Theoretical Issues in Ergonomics Science*, *21*(4), 369–396. https://doi.org/10.1080/1463922X.2018.1547459

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. https://doi.org/10.1177/0018720814539505

McCarley, J. S., & Wickens, C. D. (2004). *Human factors concerns in UAV flight*. Technical Report AHFD-05-05/FAA-05-01. Aviation Human Factors Division Institute of Aviation, University of Illinois. Savoy, Illinois, USA.

McKendrick, R., & Harwood, A. (2019). Cognitive workload and workload transitions elicit curvilinear hemodynamics during spatial working memory. *Frontiers in Human Neuroscience*, *13*(405), 1–16. https://doi.org/10.3389/fnhum.2019.00405

Morgan, J. F., & Hancock, P. A. (2011). The effect of prior task loading on mental workload: An example of hysteresis in driving. *Human Factors*, *53*(1), 75–86. https://doi.org/10.1177/0018720810393505

Moss, M., Good, V. S., Gozal, D., Kleinpell, R., & Sessler, C. N. (2016). A critical care societies collaborative statement: Burnout syndrome in critical care health-care professionals a call for action. *American Journal of Respiratory and Critical Care Medicine*, *194*(1), 106–113.

https://doi.org/10.1164/rccm.201604-0708ST

Mracek, D. L., Arsenault, M. L., Day, E. A., Hardy, J. H., & Terry, R. A. (2014). A multilevel approach to relating subjective workload to performance after shifts in task demand. *Human Factors*, *56*(8), 1401–1413. https://doi.org/10.1177/0018720814533964

Ouma, J. A., Col, L., Wayne Chappelle, F. L., & Amber Salinas, A. (2011). *Facets of occupational burnout among U.S. Air Force active duty and National Guard/Reserve MQ-1 Predator and MQ-9 Reaper operators*. USAF School of Aerospace Medicine. http://www.dtic.mil

Rothrock, L., Koubek, R., Fuchs, F., & Haas, M. (2002). Reappraisal of adaptive interfaces: Toward biologically inspired paradigms. *Theoretical Issues in Ergonomics Science*, *3*(1), 47–84. https://doi.org/10.1080/14639220110110342

Salfinger, A., Retschitzegger, W., & Schwinger, W. (2013). Maintaining situation awareness over time -- A survey on the evolution support of situation awareness systems. *2013 Conference on Technologies and Applications of Artificial Intelligence*, 274–281. https://doi.org/10.1109/.60

Sebok, A., Wickens, C. D., & Sargent, R. (2015). Development, testing, and validation of a model-based tool to predict operator responses in unexpected workload transitions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 622–626. https://doi.org/10.1177/1541931215591136

Sibley, C., Coyne, J., & Morrison, J. (2015). Research considerations for managing future unmanned systems. *AAAI Spring Symposium*, 54–59.

Sibley, C. M., Foroughi, C. K., Olson, T., Moclaire, C., & Coyne, J. T. (2017). Practical considerations for low-cost eye tracking: An analysis of data loss and presentation of a solution. In D. Schmorrow & C. Fidopiastis (Eds.), *International Conference on Augmented Cognition*, 236–250. https://doi.org/10.1007/978-3-319-58628-1

Szalma, J. L. (2009). Individual differences in human–technology interaction: Incorporating variation in human characteristics into human factors and ergonomics research and design. *Theoretical Issues in Ergonomics Science*, *10*(5), 381–397. https://doi.org/10.1080/14639220902893613

van Wulfften Palthe, O. D. R., Neuhaus, V., Janssen, S. J., Guitton, T. G., & Ring, D. (2016). Among musculoskeletal surgeons, job dissatisfaction is associated with burnout. *Clinical Orthopaedics and Related Research*, *474*(8), 1857–1863. https://doi.org/10.1007/s11999-016-4848-6

Vine, S. J., Masters, R. S. W., McGrath, J. S., Bright, E., & Wilson, M. R. (2012). Cheating experience: Guiding novices to adopt the gaze strategies of experts expedites the learning of technical laparoscopic skills. *Surgery (United States)*, *152*(1), 32–40. https://doi.org/10.1016/j.surg.2012.02.002

Wickens, C. D. (1980). The structure of attentional resources. In R. S. Nickerson (Ed.), *Attention and Performance VIII* (pp. 239–257). Psychology Press.

Wickens, C. D. (2002). Situation awareness and workload in aviation. *Current Directions in Psychological Science*, *11*(4), 128–133. https://doi.org/10.1111/1467-8721.00184

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*(3), 449–455. https://doi.org/10.1518/001872008X288394.

Wickens, C. D., & Gutzwiller, R. S. (2017). The status of the strategic task overload model (STOM) for predicting multi-task management. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 757–761. https://doi.org/10.1177/1541931213601674

Wickens, C. D., Santamaria, A., & Sebok, A. (2013). A computational model of task overload management and task switching. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 763–767. https://doi.org/10.1177/1541931213571167

Williams, K. W. (2006). *Human factors implications of unmanned aircraft accidents: Flight-control problems*. https://doi.org/10.1016/S1479-3601(05)07008-6

Young, M. S., & Stanton, N. A. (2002). Malleable attentional resources theory: A new explanation for the effects of mental underload on performance. *Human Factors*, *44*(3), 365–375. https://doi.org/10.1518/0018720024497709