The Impact of Automation Reliability and Fatigue on Reliance

Ryan Wohleber
University of Central Florida

THE IMPACT OF AUTOMATION RELIABILITY AND FATIGUE ON RELIANCE

by

RYAN W. WOHLEBER
M.S. University of Central Florida, 2015
M.A. University of Cincinnati, 2012
B.S. University of Cincinnati, 2009

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Modeling & Simulation
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Spring Term
2016

Major Professor: Gerald Matthews

ABSTRACT

The objective of this research is to inform the design of dynamic interfaces to optimize unmanned aerial vehicle (UAV) operator reliance on automation. A broad goal of the U.S. military is to improve the ratio of UAV operators to UAVs controlled. Accomplishing this goal requires the use of automation; however, the benefits of automation are jeopardized without appropriate operator reliance. To improve reliance on automation, this effort sought to accomplish several objectives organized into phases. The first phase aimed to validate metrics that could be used to gauge operator fatigue online, to understand how the reliability of automated systems influences subjective and objective responses, and to understand how the impact of automation reliability changes with different levels of fatigue. To that end, this study employed a multiple UAV simulation containing several tasks. Findings for a challenging Image Analysis task indicated a decrease in accuracy and reliance with time. Both accuracy and reliance were lower with an unreliable automated decision making aid (60% reliability) than with a reliable automated decision making aid (86.7% reliability). Further, a significant interaction indicated that reliance diminished more quickly when the automated aid was less reliable. Concerning the identification of possible eye tracking measures for fatigue, metrics for percentage of eye closure (PERCLOS), blinks, fixations, and dwell time registered changes with time on task. Fixation metrics registered reliability differences. The second phase sought to use outcomes from the first phase to build two algorithms, based on eye tracking, to drive continuous diagnostic monitoring, one simple and another complex. These algorithms were intended to diagnose the passive fatigue state of UAV operators and used subjective task engagement as the dependent variable. The simple algorithm used PERCLOS and total dwell time within the automated tasking area. The complex algorithm added percent of cognitive fixations and frequency of express fixations. The complex algorithm successfully predicted task engagement, primarily on the strength of percentage of cognitive fixations and express fixation frequency metrics.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Air Force Office of Scientific Research (AFOSR) in collaboration with Air Force Research Laboratory (AFRL). The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of AFRL or the U.S. Government.

This research would not have been possible without the help, advice and support of many. I wish to thank the members of my committee: Gerald Matthews, Ph.D., for serving as my primary advisor through my graduate career, Lauren Reinerman-Jones, Ph.D., Gregory Funke, Ph.D., James Szalma, Ph.D., and Florian Jentsch, Ph.D., for providing the opportunity for me to work on this project. I am also grateful to Air Force Office of Scientific Research scientists for their sponsorship and support in realizing the project including Gloria Calhoun, Ph.D. and Heath Ruff. I am thankful to AFRL scientists, in addition to Gregory Funke, Ph.D. who served as a committee member, including Allen Dukes for processing the eye tracking data using AFRL software and computing power and Michael Tolston, Ph.D. for his work on the recurrence quantification analysis. I would also like to thank my colleagues at the Institute for Simulation and Training including Jinchao Lin for his insight, project support, and data collection efforts, Jonathan Harris and Daniel Barber, Ph.D. for their technical expertise and troubleshooting. I would like to thank the RAs who have helped with data collection and study management including Angela Kestner, Lisa Wick, Ysabele Barrios, and Rebekah Jones. I would like to thank my friends and past colleagues, including Bee Nash. Finally, I would like to thank Bethany Carson, my parents Amy and Bill, my brothers Cameron, Daniel and Kyle, sister-in-law Jen, nephew Jacob, and Goddaughter Carolyn, for their love, support, and encouragement throughout this process.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Unmanned aerial vehicles (UAVs) have been developed and operated by the U.S. Military since 1917 (Gertler, 2012). More recently, research and development efforts have intensified as new technologies have rendered UAVs more practical and effective for several civilian (Coifman, McCord, Mishalani, & Redmill, 2004; Rango et al., 2006) and military applications (Bone & Bolkcom, 2003; Haddal & Gertler, 2010; Haulman, 2003). UAVs offer numerous advantages to military operations, notably force multiplication (Chappelle, McDonald, & King, 2010) and elimination of risk to operator's lives (Gertler, 2012; Stulberg, 2007). Whereas conventional aircraft are limited by the endurance of on-board pilots, UAV operators can work in shifts to support prolonged operations (Barnes & Matz, 1998; Tvaryanas et al., 2006). This makes them especially useful for intelligence, surveillance, and reconnaissance (ISR) missions (Chappelle et al., 2010; Gertler, 2012; Haulman, 2003).

An important development in UAV operation is the move from teams of two or more operators controlling a single UAV to multi-aircraft control (MAC) by a single team or pilot (Calhoun, Ruff, Draper, & Wright, 2011; Donmez, Cummings, & Graham, 2009; Eggers & Draper, 2006; Johnson, Leen, & Goldberg, 2007). MAC brings with it a number of new human factors challenges. Operators have limited attentional resources to be shared between the multiple tasks required to operate several UAV; it is necessary to use automation to support operators (Cummings, Clare, & Hart, 2010; Mouloua, Gilson, & Hancock, 2003; Schulte & Donath, 2011). Automation promises to increase operator efficiency, improve safety, enhance the flexibility of operations, and lower operator workload (Cummings, Brzezinski, & Lee, 2007; Dahm, 2010; Dixon, Wickens, & Chang, 2005; Parasuraman & Manzey, 2010). Unfortunately these benefits are jeopardized without optimal reliance on automation (Parasuraman & Manzey, 2010). To properly implement automation, it is important to understand what types and levels of automation best support operators and their

mitigation strategies (Eggers & Draper, 2006; Endsley & Kaber, 1999). Further, it is important to understand the implications and challenges of the new supervisory role imposed on operators by the proliferation of automation (Dixon, Wickens, & McCarley, 2007; Parasuraman & Riley, 1997; Ruff, Calhoun, Draper, Fontejon, & Guilfoos, 2004; Sheridan & Parasuraman, 2005).

UAV operation is marked by long periods of low workload (Cummings, Mastracchio, Thornburg, & Mkrtchyan, 2013) interspersed with brief periods of intense activity (Cummings et al., 2007). Even with the aid of automation, attentional demands of multiple UAV operation during periods of high activity is daunting and operator reliance on automation must be must be optimal to deal effectively with these events (Cummings et al., 2010). Conversely, automation provides opportunities for operators to disengage during periods of low workload when the use of automation may be detrimental to an operators' ability to maintain vigilance (Bailey & Scerbo, 2007; Cummings et al., 2013). It is, therefore, necessary to optimize operators' reliance on automation (Calhoun et al., 2013, 2011; Endsley & Kaber, 1999).

There are several factors that can contribute to operator reliance on automation including individual differences (Chen & Barnes, 2012; Chen & Terrence, 2009; Lin et al., 2015; Szalma & Taylor, 2011), training (Drury, Richer, Rackliffe, & Goodrich, 2006; Helmreich, Merritt, & Wilhelm, 1999; Parasuraman & Manzey, 2010; Pavlas et al., 2009), operator fatigue (Neubauer, Matthews, Langheim, & Saxby, 2012; Parasuraman & Riley, 1997), interface design (Parasuraman & Manzey, 2010), reliability of automation (Chen & Barnes, 2012; Rovira, McGarry, & Parasuraman, 2007), and level of automation (Calhoun et al., 2013, 2011; Johnson et al., 2007). The present effort is primarily concerned with the impact of operator fatigue, automation reliability, and level of automation on reliance. In automation assisted operations it is important to keep operators "in the loop" and in ultimate control of the activities of all vehicles operated to guard against unintended design consequences of full automation and imperfect reliability (Billings, 1991; Hanson & Harper, 2000;

Miller & Parasuraman, 2003; Ruff, Narayanan, & Draper, 2002). Intermediate levels of automation often place operators in a supervisory role where the machine makes decisions by default and the operator monitors the automation, able to override automated decisions if necessary. Previous work has articulated concerns that placing operators in such a role is apt to induce passive fatigue (Parasuraman, Sheridan, & Wickens, 2000; Ruff & Calhoun, 2011; Sauer, Kao, & Wastell, 2012; Warm, Matthews, & Finomore, 2008). Passive fatigue can cause operators to take an energy conservation strategy to tasking (Sauer, Wastell, Robert, Hockey, & Earle, 2003) which may manifest as overreliance on automation. Neubauer and colleagues (Neubauer et al., 2012; Neubauer, Matthews, Saxby, & Langheim, 2011) found that fatigue was associated with greater reliance on full automation, even when reliance did not result in improved performance. A fitting use for automation in UAV tasks is to place the operator in an intermediate role, where automation aids in decision making to some degree, but can be overridden. There remains a need to examine the impact of fatigue on the use of intermediate levels of automation such as automated decision-making aids in UAV operations.

Reliability of automation also has important implications for reliance on automation; because automation is almost never perfectly reliable, investigations of reliance must account for imperfect automation (Dixon & Wickens, 2006). Previous research has shown that reliability of automation predicts operator trust in automation (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Lee & See, 2004; Parasuraman, Molloy, & Singh, 1993) which in turn predicts reliance on automation (Merritt & Ilgen, 2008; Parasuraman & Riley, 1997). When serving in a monitoring or backup role, reliable systems tend to induce overtrust in operators. Overtrust can result in complacency or misuse of automation and a corresponding loss of situation awareness and task performance (Parasuraman et al., 1993; Parasuraman & Wickens, 2008). This tendency is exacerbated by fatigue and more likely when operators are responsible for multiple tasks. Conversely, unreliable systems lead to distrust and

resulting disuse resulting in higher operator workload, subsequent task shedding, and task performance decrements (Parasuraman & Riley, 1997). Thus, to achieve optimal reliance it is important to properly calibrate trust in automation (Lee & See, 2004; Merritt, Lee, Unnerstall, & Huber, 2014; Parasuraman & Miller, 2004).

There is a long history of research based on the belief that adaptive interfaces can improve the balance between task, machine, and human needs (Hancock & Chignell, 1988). More recent work has investigated using adaptive level of automation (LOA) to promote proper reliance (Parasuraman & Wickens, 2008). To calibrate trust properly, it is necessary to understand not just the likely impact of reliability on trust in automation, but also how the impact of reliability changes with level of fatigue. The present effort aims to use an understanding of these variables to inform interventions, which will adapt the level of automation in a system online to promote proper reliance on automation.

Fatigue

Numerous reports, studies, and reviews attest to the operational significance of fatigue (Chappelle, Salinas, & McDonald, 2011; Ouma, Chappelle, & Salinas, 2011; Tvaryanas et al., 2006; Tvaryanas & Macpherson, 2009; Tvaryanas & Thompson, 2006). Indeed, reports have indicated that fatigue is greater for UAV operators than for manned aircraft operators (Chappelle, Salinas, et al., 2011; Tvaryanas et al., 2006; Tvaryanas & Thompson, 2006). Several factors associated with task demands have been identified as causes of fatigue in UAV operation including long shift duration and human-machine interface difficulties (Chappelle, Salinas, et al., 2011; Ouma et al., 2011). Tvaryanas (2006), for example, identified several human-machine interface challenges faced by UAV crews using a ground control station (GCS) including relative sensory deprivation (e.g. no auditory or haptic cueing) and the need to depend almost exclusively on focal vision (central 30% of visual field) to determine vehicle state. Although factors that induce fatigue in RPV operation have been

identified and studied, more work is necessary to understand how fatigue might be managed by task characteristics in UAV operation.

The following discussion will summarize relevant literature on fatigue and touch on topics important for the goal of optimizing operator reliance on automation. Although it will be necessary at times to speak about fatigue in relation to automation, I will refrain from any substantial discussion of automation until the trust and reliance section.

Theory and Implications

Despite an extensive literature on the construct, there has yet to be agreement on a precise and consistent definition of fatigue (Matthews, Desmond, Neubauer, & Hancock, 2012) or consensus on how fatigue should be measured (Christodoulou, 2012). Instead, there exists numerous definitions, each colored by the context of a problem to be addressed or the theoretical perspective of the investigator (Christodoulou, 2012). Desmond and Hancock (2001) explain that part of the difficulty with defining fatigue has to do with the inherent circularity of defining a construct. To precisely define a construct, one must be able to measure a phenomenon reliably and validly, however one needs a precise definition of the construct in order to develop a reliable and valid measure (Muscio, 1921). Instead, definitions of fatigue stem from personal experience of the reality of fatigue, which results in considerable difficulty. This difficulty with finding a suitable definition for fatigue is exacerbated by the fact that fatigue is complex, multifaceted, and as discussed below, almost certainly multidimensional (Hancock & Desmond, 2001; Matthews, Desmond, Neubauer, et al., 2012).

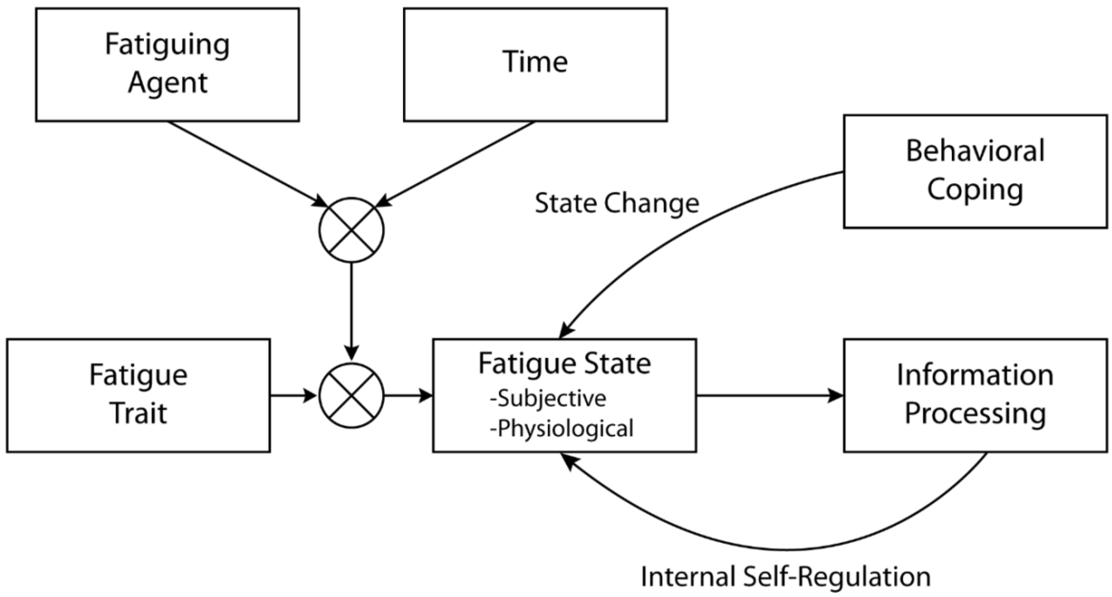


Figure 1. A simple trait-state model for fatigue. Adapted from Matthews, Desmond, & Hitchcock, 2012.

To understand and distinguish the relevant facets of fatigue, a model such as the example in Figure 1, adapted from Matthews, Desmond, and Hitchcock (2012), is useful. This model distinguishes between trait and state fatigue. Trait fatigue refers to proneness to experience fatigue or vulnerability to factors that induce the experience. State can be defined as an temporary quality that pervades consciousness driven by underlying mental processes which generalize across contexts (Matthews et al., 2002, p. 316). Specifically, time-dependent external agents moderate the influence of fatigue trait on fatigue state. External agents that could make trait fatigue more apparent might include prolonged work period (Tvaryanas et al., 2006) or low levels of sleep (Hockey, Wastell, & Sauer, 1998). Time influences might be long term, like circadian rhythm (Banks, Jackson, & Van Dongen, 2012), short term like time performing a task (e.g. Saxby et al., 2008), or both (e.g. Sauer et al., 2003). Fatigue state, in turn, influences information processing and behavioral coping.

Fatigue state can be expressed behaviorally, physiologically, and subjectively (Matthews, Desmond, & Hitchcock, 2012). The interrelationships between these different expressions of fatigue

vary with context and individual differences. Consequentially, relationships with high face validity are often unreliable. For example, subjective fatigue does not always predict performance decrements (Thorndike, 1914). With these points in mind, it is important to focus on aspects of fatigue theory that are applicable specifically to the context of UAV operation to understand the impact of fatigue on UAV operators.

Broadly, fatigue can be measured using dimensions, similar to the way personality is often measured along five dimension using the Five Factor Model (Matthews, Desmond, & Hitchcock, 2012; McCrae & Costa, Jr., 2008). An advantage of this approach is the generalizability of the dimensions to different domains. Fatigue can be and has been assessed along a single dimension. Assessment of some prominent fatigue scales (Ahsberg, Gamberale, & Gustafsson, 2000; Lai et al., 2011), including scales that are meant to be multidimensional (De Vries, Michielsen, & Van Heck, 2003), has lent support to a single dimension conceptualization (Christodoulou, 2012). De Vries et al. (2003) found that when four different fatigue scales were combined, including two design to assess fatigue as a multidimensional construct, a single factor accounted for 60 percent of the variance.

Despite the utility of a unidimensional conceptualization, Matthews, Desmond, and Hitchcock (2012) point out that support for unidimensional conceptualization based on the analysis of several fatigue scales may be the result of unbalanced item sampling. They point out for example, that the fatigue surveys in an analysis conducted by Michielsen, De Vries, and Van Heck (2003) all have subscales for tiredness but don't overlap as much with other aspects of fatigue. This resulting overemphasis on tiredness items may have created artificial unidimensionality. With a broader scope, fatigue seems to be multidimensional and it is important to account for these multiple dimensions because different facets may vary in their impact on performance (Matthews, Hancock, & Desmond, 2012). Further, facets of fatigue overlap with stress, which is also multifaceted (Matthews, Szalma,

Panganiban, Neubauer, & Warm, 2013). Surveys of RPV operators for example identified long work hours as a source of occupational stress (Ouma et al., 2011). Vigilance work can be stressful when combined with a need to maintain high alertness (Thackray, 1980). This stress in turn can magnify the effects of fatigue.

With the overlapping importance of stress and fatigue in mind, this effort followed the comprehensive subjective state model, developed by Matthews et al. (2002), which accounts for both constructs and integrates established theoretical domains dealing with self-regulation in task performance (e.g., energy and tension, Thayer, 1989; Watson & Clark, 1984; and cognitive interference, Sarason, Sarason, Keefe, Hayes, & Shearin, 1986; Wells & Matthews, 2015). Multidimensional models of fatigue have traditionally differentiated between affective, motivational, and cognitive components of fatigue, among others (Hilgard, 1980; Matthews, Desmond, & Hitchcock, 2012; Mayer, Chabot, & Carlsmith, 1997). Matthews et al. (1999) employed test items from these components of subjective experience in an attempt to elucidate higher order constructs of subjective state in a task performance setting. Although traditionally thought of as separate components (Mayer et al., 1997), evidence compiled by Matthews and colleagues (c.f., Matthews et al., 2002, 2013) suggest that affect, motivation, and cognition function together to support higher order processes. Their model organizes these higher order processes into three fundamental state dimensions: task engagement (energy, concentration, task motivation), distress (negative mood, lack of perceived control) and worry (intrusive, distracting thoughts).

Task engagement and distress integrate affective, cognitive, and motivational, domains, whereas worry is purely cognitive. Theoretically, states measured along these parameters may reflect different modes of self-regulation (Matthews et al., 2002), which can be described as sets of processes and behaviors sustaining the pursuit of personal goals within a fluctuating external environment (Matthews, Schwean, Campbell, Saklofske, & Mohamed, 2000). Task engagement is of

particular interest in the current effort as it parallels Hockey's (1997) fatigued mode of cognitive control, which predicts adaptive shifts toward low effort modes of performance after prolonged work. Mental fatigue is typically registered as a large-magnitude decrease in task engagement (Matthews et al., 2013; e.g., Matthews, Warm, Reinerman-Jones, et al., 2010; Teo & Szalma, 2011). Task engagement has also been associated with physical fatigue. Notably, Matthews et al. (2002) found that task disengagement was especially prevalent in visual vigilance tasking relative to auditory vigilance and working memory tasks. This outcome parallels previous findings of eyestrain as a source of fatigue in vigilance tasks (Temple et al., 2000) and is an important consideration in the context of UAV operation, which requires prolonged use of focal vision (Tvaryanas, 2006).

In research on fatigue and performance, use of a multidimensional model is imperative because difference types of task demand elicit different patterns of response. (Matthews et al., 2013). Desmond and Hancock (2001) outlined a model of fatigue which makes a distinction between active and passive fatigue states. Active fatigue is characterized by continued and prolonged perceptual-motor adjustment to tasks. That is, active fatigue can result from situations in which successful ongoing task performance is contingent on frequent perceptual sampling of the environment and behavioral adjustment. As active fatigue increases with time on task, sampling rate decreases and performance adjustments are less frequent but greater in magnitude to account for greater deviations from task goals resulting from lower sampling rate. An example of a situation that may induce active fatigue is driving on a highway with strong crosswinds (see Saxby et al., 2008; Saxby, Matthews, Hitchcock, & Warm, 2007). Here, a driver would need to assess continuously the trajectory of the vehicle and make frequent course corrections to maintain course. Passive fatigue, in contrast, is elicited by tasks requiring rare perceptual-motor response where fatigue stems from chronic under-stimulation (Desmond & Hancock, 2001). Sampling of the environment decreases over time as an

adaptive response to the infrequent need to respond. Monitoring tasks are typically associated with this passive fatigue.

Recent studies of driving fatigue (Neubauer et al., 2011; Saxby et al., 2008, 2007; Saxby, Matthews, Warm, Hitchcock, & Neubauer, 2013) have confirmed that active and passive forms associate with different patterns of subjective state response. In this work, passive fatigue manipulations (monitoring an automated drive) resulted in a precipitous loss of task engagement and an accompanying loss of alertness, operationalized as response time to a critical event. Conversely, active fatigue manipulations had a modest impact on task engagement, but also generated large and stable increases in distress. An important caveat is that distress may be more susceptible to the short-term impact of events while performing a task. In the 2008 effort (Saxby et al., 2008), an emergency event following a driving automation monitoring task (the passive fatigue manipulation) elevated distress to the level of that in the active fatigue condition. Nevertheless, the emergency nature of the distressing event did not eliminate the response time decrements in the passive fatigue condition. In line with Desmond and Hancock's (2001) theory, passive fatigue was associated with low workload and active fatigue with high workload.

Active and passive fatigue can be linked to UAV operation. Broadly, UAV operation is characterized by long periods of low workload interspersed with shorter periods of high mental workload (Mouloua, Gilson, Kring, & Hancock, 2001), which may, in accordance with Desmond and Hancock's (2001) theory, induce passive and active fatigue respectively. Task characteristics that may induce passive fatigue include the relative sensory deprivation of being separated from the environment in which the vehicles operate (Van Erp, 2000), as well as the passive role of monitoring highly automated flight management systems (Mouloua et al., 2001). In this isolated supervisory role, it is not uncommon for operators to spend large amounts of time waiting for a critical event to

occur (Cummings et al., 2013). Together, these context and task characteristics fit the description of a situation likely to induce passive fatigue.

The monitoring task characteristics of UAV control just described require vigilance. Vigilance refers to the ability to maintain attention on task and alertness to stimuli over a prolonged period (Warm, Parasuraman, & Matthews, 2008). "Vigilance decrements" refers to declining performance with time on task, which occurs with vigilance tasks. This decrement usually appears within the first 15 minutes of the task, and is robust to experience and context. The diversity of contexts under which vigilance occurs has been explained by attentional resource theory (Mackie, Parasuraman, & Davies, 1977; Warm & Dember, 1998). This theory suggests that vigilance decrements occur when information-processing resources are depleted faster than they can be replenished.

This conceptualization suggests that vigilance is underpinned by a general process factor, however. Parasuraman (1979) refined this conception based on performance similarities and transfer of training within but not between two types of target detection: successive and simultaneous (Parasuraman, 1976). Successive detection, where stimuli appear one after the other, requires discrimination comparison of a current stimulus with memory of a critical signal. Simultaneous discrimination presents critical and non-critical signals together for direct comparison. Task characteristics, such as spatial uncertainty of the occurrence of critical signals and the level of multitasking influence performance on successive tasks to a greater degree, and successive vigilance tasks are generally induce higher workload. Thus, although vigilance decrements are not context specific, they are task type specific. Type specificity notwithstanding, vigilance is mentally demanding and stressful (Warm, Parasuraman, et al., 2008). Several studies have demonstrated that time on vigilance tasks decreases task engagement and increases distress (Szalma et al., 2004; Warm,

Matthews, et al., 2008). Thus, the vigilance element of operations may be particularly at risk because vigilance tasks are often a source of cognitive fatigue (Hockey, 1997).

In addition to the fatiguing influence of vigilance tasks, extensive literature on signal detection (Warm, Parasuraman, et al., 2008) has shown that vigilance task performance itself is vulnerable to fatigue. Matthews and colleagues demonstrated that even minor levels of pre-task fatigue predicts reduced perceptual sensitivity on tasks requiring sustained attention (Matthews, Warm, Reinerman-Jones, et al., 2010; Matthews, Warm, Reinerman, Langheim, & Saxby, 2010; Shaw et al., 2010).

Accordingly, the symptoms of current UAV operation correspond to those characteristic of passive fatigue and vigilance tasking. Reports of UAV operation attest that the infrequency of signals and lack of interaction make it difficult for operators to maintain alertness and predispose them to "hazardous states of awareness" (Pope & Bogart, 1992; Tvaryanas, 2006; Tvaryanas et al., 2006). UAV operators working continuous shifts report boredom, and their performance shows significant loss of sustained attention and vigilance, decreased performance accuracy, and slowed response times (Chappelle et al., 2010; Cummings et al., 2013; Ouma et al., 2011; Tvaryanas et al., 2006).

Active fatigue factors are not as well documented in relation to current UAV operation. Walters, French, and Barnes (2000) found that reduced UAV crew size negatively impacts performance in a computer model of sleep deprivation and fatigue. However, the impact of sustained high workload operation in healthy populations is unclear. The lack of sustained high workload studies in UAV operation may be a result of staffing practices, which guards against overload and minimizes the occurrence of situations that demand prolonged high workload. Current UAV operations utilize two or more operators to pilot and operate sensors for a single UAV (Dixon, Wickens, & Chang, 2004; Ouma et al., 2011). The substantial body of research on passive fatigue in current operation, in contrast, may reflect that a satisfactory solution to the passive fatigue

has yet to be found. Although active fatigue may not be a pressing contemporary problem, it is expected to become an critical issue as the ratio of operators to UAVs decreases and single operators are responsible for several highly automated UAVs (Cummings et al., 2007; Lewis et al., 2010; Schulte & Donath, 2011).

Risk of active fatigue in single operator management of multiple UAVs stems not just from a general increase in task demand, but also from the multi-tasking nature of UAV operation (Ouma et al., 2011; Schulte & Donath, 2011). Single operators will need to assume disparate roles currently filled by teams of at least two operators. Current MQ-1 Predator and MQ-9 Reaper missions currently require a pilot and a sensor operator. Pilot tasking includes intermittently maneuvering the UAV manually for strategic positioning and to avoid bad weather (despite automated flight capability for more routine flight); performing in-flight mission planning in accordance with theater rules of engagement (ROE); receiving, interpreting, and executing orders and instructions; monitoring systems controls; coordinating with air traffic control, ground crew, and other aircrew; and more (see Chappelle, McDonald, & McMillan, 2011 for an exhaustive listing). These tasks involve attending to and interpreting information from several visual and auditory sources and maintaining awareness of the state of several systems (Ouma et al., 2011). Sensor operators duties include finding, tracking, and monitoring airborne, maritime, and ground targets; discriminating between valid and invalid targets; receiving and acting on target briefs; conducting battle damage assessments; providing target intelligence to command and aircrew; and operating laser target marking systems and weapons delivery systems. Like pilots, sensor operators must maintain vigilance of multiple systems and receive and interpret various types of information simultaneously, however sensor operator tasking is particularly demanding in regards to the discrimination and synthesis of visual information (Ouma et al., 2011).

The multiple roles and duties a single operator must take up to control multiple UAVs present a considerable challenge. Indeed, Mouloua et al. (2003) asserted that the dual concerns of mission requirements and flight control may even prohibit single operator control of multiple UAVs. Monitoring multiple UAVs may subject operators to extended periods of high workload, requiring extensive multi-tasking (J. M. Riley & Endsley, 2005). Multitasking entails strategic management of several tasks and their components. According to Hockey's (1997) regulatory control model, the maintenance of performance stability in challenging conditions is an active and effortful process. This active control allows an operator to regulate the effectiveness of task behaviors in relation to competing goals, changing demands, and current levels of energetic resources. Since multitasking entails active control, work that demands extensive multitasking is especially vulnerable to effort-minimization strategies, including task shedding (Schulte & Donath, 2011), and switching from a proactive to a reactive control mode (Hockey et al., 1998; Sauer et al., 2003). The risk of effort minimization may be exacerbated by fatigue; after prolonged work, the infrequency of critical signals in monitoring tasks may not be enough to motivate fatigued operators to continue the effortful strategies required to maintain adequate supervision of automated systems.

Some initial work on the characteristics of task related fatigue in a multiple UAV environment was conducted at the University of Cincinnati (UC). Guznov, Matthews, Funke, and Dukes (2011) found that a simulated multiple robot control task had a propensity to produce the high levels of workload and distress typical of active fatigue. Subsequent work reproduced these findings and found additionally that task engagement declined over time in single operators but not in teamed operators (Matthews et al., 2011), which is an important consideration in the move from teams of operators to solo operation of multiple UAVs. Guznov (2011) found that task engagement generally declined in a UAV task which required the detection of visual ground targets. This decline was especially pronounced when operators were not required to perform a secondary task of

responding to audio messages, which suggests that multi-tasking may in fact, guard against the effects of passive fatigue. The large magnitude decline of task engagement in the single task condition resembled the passive fatigue response Saxby and colleagues (Saxby et al., 2008, 2007) registered in a simulated drive which required participants to monitor a fully automated drive. Guznov (2011) confirmed this finding and additionally found that pre-task fatigue predicted detection performance, supporting the interactive conception of fatigue as influence and consequence of performance.

More recent work (Panganiban, 2013; Panganiban & Matthews, 2014; Wohleber, Matthews, Reinerman-Jones, Panganiban, & Scribner, 2015) compared the impact of several workload manipulations on stress and fatigue in a multitasking environment using the Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles (RESCHU) simulation (Donmez, Nehme, & Cummings, 2010). This simulator required participants to steer multiple UAVs to appropriate target locations while avoiding hazards and simultaneously perform visual searches for target objects in a sensor window. A high task-load condition, which, relative to a low task-load condition, required management of more RPVs under greater time pressure, produced increased operator task engagement and distress. These findings are similar to Guznov's and colleagues (Guznov et al., 2011) findings using the multi-RPV RoboFlag simulator environment and consistent with symptoms of active fatigue. Additionally an evaluative stress condition, which provided participants with scripted negative feedback, diminished task engagement, and increased distress beyond that induced by high task load. Task variables were not manipulated to induce passive fatigue; duration of each UAV simulation trial was only 10 to 15 minutes.

A large body of work on fatigue in air traffic control (ATC) is also informative inasmuch as ATC task characteristics are similar to those anticipated for multi-UAV operation (see Backs, Navidzadeh, & Xu, 2000; Brookings, Wilson, & Swain, 1996; Desmond & Hoyes, 1996; Desmond &

Matthews, 2009; Endsley, 1988; Endsley & Rodgers, 1997; Gander, 2001; Hurter, Serrurier, Alonso, Tabart, & Vinot, 2010; Langan-Fox, Sankey, & Canty, 2009; Metzger & Parasuraman, 2001; Rosekind et al., 1996a, 1996b, 1997; Smolensky, 1990; Straussberger, 1996; Vanderhaegen, Crevits, Debernard, & Millot, 1994; Wilson & Russell, 2003). Useful parallels have been drawn between fatigue issues for supervisory roles in UAV operation and for air traffic control (e.g., Cummings et al., 2013; Dixon et al., 2005; Donmez et al., 2010; Drury, Riek, & Rackliffe, 2006; Sheridan & Parasuraman, 2005). Work on fatigue in ATC has confirmed breakdowns in monitoring as a result of prolonged periods of low workload (Cummings et al., 2013; Straussberger, 1996; Thackray, 1981) characteristic of passive fatigue. Active fatigue, however, does not appear to be a well-documented phenomenon in ATC despite the fact that use of high workload manipulations in simulated ATC tasking is fairly common in laboratory settings. Although Della Rocco (1999) found that a substantial proportion of air traffic control incidents report workload as the cause for operator error, Melton (1982) argues, based on physiological studies of air traffic controllers at work, that modern ATC in the field is not unusually stressful and suggested that accounts of controller stress is exceptional rather than typical. Hale, Williams, Smith, and Melton (1971) for example, found that ATCS's stress levels were comparable to those of off-duty laboratory scientists. A greater focus on passive fatigue issues in the implementation of automation in multi-UAV operation may also be appropriate.

Research on the implementation of automation in ATC typically focuses on proper calibration of task-load to maintain operator performance (Sarter, Woods, & Billings, 1997; Thackray, 1980). Like ATC, it is anticipated that highly automated, multiple-UAV operations will consist of both periods of high task demand and periods of little or no operator engagement (Eggers & Draper, 2006; Sarter et al., 1997). Laboratory work has demonstrated that abrupt transitions between periods of low and high demand increase distress in operators, whether the transition is

anticipated or not (Helton, Shaw, Warm, Matthews, & Hancock, 2008). Earlier field studies have indicated, however, that transitions are gradual and allow air traffic controllers to adapt effectively to changes in demand (Sperandio, 1978). It is unclear how abruptly task load shifts may occur in multi-UAV operation, but it is likely that sensor operation tasking, which may not map to ATC, may influence the quickness of workload shifts. In the event that the multiple roles of an operator call for rapid shifts in task-load, active fatigue may result from the need to regulate task performance strategies continuously.

A final similarity between ATCS's and multi-UAV control is that operators occupy a fixed space away from the physical location of the aircraft and must depend on displays for situation awareness (SA; Drury, Riek, et al., 2006). SA has several definitions, but a commonly accepted definition (e.g., Adams, 2007) provided by Endsley defines it as:

> "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" (Endsley, 1988)

Maintaining SA is important for UAV operation performance (Barnes & Matz, 1998; Drury, Riek, et al., 2006; J. M. Riley & Endsley, 2005) but it can be affected by both active (Chen & Barnes, 2012) and passive fatigue (Ruff et al., 2004) conditions . With active fatigue, SA is compromised by a task shedding coping strategy, which occurs when the operator aims to maintain performance on the primary task but can no longer allocate resources to other tasks. Conversely, increased automation, which can induce passive fatigue (discussed further below), has been linked to decreased SA (Calhoun et al., 2011; Parasuraman et al., 2000). Operators are less aware of changes in the environment when automation or another operator are in charge of those changes (Endsley, 1996; Endsley & Kiris, 1995).

Measurement of Fatigue

Fatigue can be assessed in several ways. This effort aimed to identify those with potential to gauge the experience of fatigue noninvasively during performance. Below I review the candidate measures that were used to this end. These measures were also used to understand the impact of fatigue on reliance on automation, discussed below.

Subjective Measures. As described above, fatigue can be understood as a state or as a trait (Matthews, Desmond, & Hitchcock, 2012). Self-report measures can be used to gauge both, but in this effort, I am concerned solely with subjective state. To this end, we used the Dundee Stress State Questionnaire (DSSQ), which gauges latent states of task engagement, distress, and worry, through their affective, motivational, and cognitive expressions. Instructions for the DSSQ emphasize responding according to immediate experience in order to capture subjective state and not trait influences (Matthews et al., 1999). This measure can be used to register the different patterns of response associated with passive and active fatigue (c.f., Lin et al., 2015; Saxby et al., 2013).

Although the DSSQ may be useful for identifying measures and metrics that relate to subjective state fatigue, methods of gauging fatigue online that do not require operator introspection must be explored for several reasons. Self-report measures are believed to provide only an indirect gauge of states, as they can only access the expressions of states into consciousness and not the state itself (Matthews et al., 2002). Both conscious and non-conscious fatigue may impact performance however (Matthews, Desmond, & Hitchcock, 2012). Driving research suggests that fatigue may impact not just the willingness of operators' to contribute more resources to a task, but also operators' awareness of performance decrements, which may serve to alert operators to the onset of fatigue (Matthews & Desmond, 2002). Several studies have documented the futility of relying on introspection for the tracking of performance decrements (Dinges, Mallis, Maislin, & Powell, IV, 1998; e.g., Brown, 1997; Dinges, 1989; Wylie, Shultz, Miller, Mitler, & Mackie, 1996). This presents a

challenge to interventions that would rely on an operators' conscious experience of fatigue state. In a recent driving study, Neubauer and colleagues (Neubauer et al., 2012) confirmed suboptimal reliance on automation as a result of fatigue; during a fatiguing drive, drivers activated automation that did not improve performance. Another challenge to subjective assessments is that cognitive heuristics involved in recall may cause the experience of fatigue to be misrepresented (Broderick et al., 2008). Therefore, subjective recall may not reliably inform an intervention. Even if subjective assessments could provide a genuine representation of subjective fatigue during performance, the administration of subjective assessments during task performance would be disruptive and intermittent, at best provide a low granularity depiction of moment-to-moment fluctuations in fatigue level.

Performance Measures. Performance measures are advantaged in that they do not require the participant to be aware of their level of fatigue (Matthews, Desmond, & Hitchcock, 2012), are not vulnerable to response bias, and provide data that can be objectively verified by an observer (Christodoulou, 2012). Performance measures have long been used to gauge fatigue and related constructs of sleepiness and drowsiness (Wang, Yang, Ren, & Zheng, 2006; Wierwille, Wreggit, Kirn, Ellsworth, & Fairbanks, 1994). In the driving domain, for example, fatigue induced by time on task has been linked to: ability of a driver to maintain lane (Dureman & BodéN, 1972), heading or position error (Desmond & Matthews, 1997; Sussman, Sugarman, & Knight, 1971), frequency of steering reversals (Matthews & Desmond, 2002; Siegmund, King, & Mumford, 1996), and the magnitude of those reversals (Sugarman & Cozad, 1972; Thiffault & Bergeron, 2003). In the latter two examples, the frequency of steering changes tends to diminish with passive fatigue conditions, and the magnitude of those reversals tend to increase. Steering reversals may also grow less frequent in active fatigue conditions (MacDonald & Hoffman, 1980). These performance outcomes can be explained in terms of the impact of fatigue on environmental sampling, discussed above (Desmond & Hancock, 2001).

In fatigue research, using a primary task to gauge subjective fatigue may be difficult due to the adoption of active coping strategies to maintain level of performance in response to fatigue (Christodoulou, 2005, 2012; Hockey, 2012; Matthews, Davies, Westerman, & Stammers, 2000). Use of an aspect of a task, such as lane keeping, may be successful because such aspects are not a salient component of the task goal and therefore not directly subject to compensatory efforts. For example, although staying within a lane while driving may be an important goal, degree of alignment with lane or frequency of corrections to maintain the lane may not themselves be important or even conscious goals. Further, automatic, unconscious processes may principally control such task components of performance. This separation of controlled and unconscious processes within a task has been demonstrated by research in domains such as visual processing (Bridgeman, 1992) and complex task training (Myers & Fisk, 1987). This conscious awareness of only portions of performance may partly account for the dissociation between subjective and performance measures and inadequacy of introspection for anticipating performance decrements produced by fatigue. The assessment of task components under automatic, non-conscious control as a gauge of fatigue is promising, but it depends on the existence of task components that are reliably and primarily under automatic control as well as the ability to measure such components.

Probe tasks are one way to measure level of fatigue by testing components not under primary control. Use of a probe in automation monitoring tasks have been demonstrated in (Metzger & Parasuraman, 2001; Neubauer et al., 2012; Saxby et al., 2008). These tasks may assess fatigue by asking the participant to recall a certain feature of the task without forewarning them about the question, or by having them react to an unexpected event. Saxby et al. (2008), for example, had participants avoid a van that suddenly appeared in the road. In studies cited here, the probe task was able to discriminate conditions that induced passive fatigue despite participants' success maintaining performance on the primary task. Unfortunately, probe tasks not ideal for providing a

continuous online gauge of fatigue level. Probes may be disruptive to task performance. In Saxby et al.'s (2008) driving fatigue study, for example, drivers were distressed following the probe such that expected differences due to the experimental manipulation were likely masked in a post-task stress state assessment. Further, less jarring probes may still be vulnerable to habituation; the operator may learn to change behavior in anticipation of them (Metzger & Parasuraman, 2001). As a result, a probe might only be used successfully one time.

Although detecting performance decrements as a result of fatigue is useful for understanding the nature of fatigue, it is more useful in operational contexts to intervene preemptively before motivation or resources to cope are exhausted and fatigue is able to impact performance on critical tasking. Therefore, instead of monitoring critical tasking, which operators must perform, an alternative solution might be to add a noncritical task to existing tasking, which could be affected by fatigue without threatening more imperative tasking. However, the addition of such a task might distract from other tasking and add to fatigue in overload situations. The system could switch the non-critical task off in such cases, but fatigue could then no longer be monitored. Finally, substantial disparities in motivation to do a non-critical task relative to critical tasking might affect performance on the non-critical task in unpredictable ways, rendering conclusions about fatigue state undependable. Other solutions are needed to prevent fatigue-related performance decrements without disrupting performance on critical tasks.

A final, important note in the discussion of performance measures for gauging fatigue is that, in this effort, I intended to investigate fatigue as a predictor of operator performance (specifically reliance). Performance of a secondary task may be an informative online gauge of the fatigue state of an operator. However, using a performance measure (e.g. a secondary task) to understand the relationship between fatigue and performance (reliance) risks circularity. Thus, a

performance-based metric is unhelpful for understanding the nature of fatigue states and their consequences.

Psychophysiological Measures. Subjective measures are limited by their inability to tap into automatic processing and therefore are an incomplete gauge of fatigue. These measures may also be difficult to administer frequently and unobtrusively during task performance. Performance measures may provide a more discrete and continuous gauge of how fatigue affects performance. However, performance based measures are not able to discriminate between fatigue and other factors that may impact performance, like task motivation. Further, performance is not an ideal gauge for preventing performance decrements because it necessarily detects fatigue only after fatigue has had some impact on performance. Psychophysiological measures promise some resolution to these shortcomings of subjective and performance based measures.

Broadbent (1971) and, more recently, others (e.g., E. R. Smith & DeCoster, 2000) have proposed that information processing is supported by two hierarchical levels, a lower level of open-loop, automatic processing, and a higher level of closed-loop, controlled processing. Whereas controlled processing is characterized by conscious and effortful processing, lower level processing functions effortlessly and largely without conscious awareness. Verwey and Zaidel (2000) hypothesized that psychophysiological measures gauge the impact of fatigue primarily at the lower, automatic level whereas subjective measures of fatigue tap into higher, controlled level processing. Psychophysiological measures may strengthen understanding of fatigue state by tapping into automatic processes, which subjective assessments cannot directly access (Craig & Tran, 2012). Further, they present a promising solution to the problems of operators' lack of awareness of fatigue decrements (Matthews & Desmond, 2002), which may keep operator control of interface parameters from being a viable solution for fatigue management (c.f., Neubauer et al., 2012). Because psychophysiological measures provide direct and continuous monitoring of fatigue state, they have

potential to detect potential threats in time to prevent decrements (Craig & Tran, 2012). Illustrating this possibility, Gevins et al. (1990) found that fatigue related changes in EEG response occurred before performance was degraded on a difficult memory and fine-motor control task.

Research has investigated fatigue and related constructs using several different psychophysiological methods including metrics derived from cardiac activity such as inter-beat interval and heart rate variability (e.g., Mascord & Heath, 1992), cerebral blood flow velocity (e.g., Matthews, Warm, Reinerman-Jones, et al., 2010), electroencephalography (EEG; e.g., Horne & Reyner, 1996), and eye tracking metrics (e.g., Verwey & Zaidel, 2000). Thus far, research exploring psychophysiological assessments to gauge fatigue in adaptive automation systems has focused on EEG metrics and event-related potential (ERP) analysis (e.g., Mikulka, Scerbo, & Freeman, 2002; Pope, Bogart, & Bartolome, 1995; Prinzel, Freeman, Scerbo, Mikulka, & Pope, 2003). With the onset of fatigue, EEG registers relatively reliable increases in slow wave activity, related to drowsiness and sleep, and Alpha wave activity, inversely related to cortical arousal (Craig & Tran, 2012). These changes in wave activity may occur before performance is impacted (noted above; Gevins et al., 1990).

Despite the promise of EEG, the present project will look at the utility of eye tracking, which is easier to set up, less invasive, can be used for detecting changes in operator interactions with automation, and is arguably a more effective gauge of fatigue online. For a UAV operator shift, the setup of a capable EEG system would require an assistant to place electrodes, a baseline calibration, and pre-task testing (e.g., impedance checking). After the shift, the system components would have to be removed, cleaned, and stored, likely with the help of an assistant. An eye tracking unit may only require a single calibration for each operator, which can be used for all subsequent shifts. Further, that calibration can be done by the operator without assistance. No post-shift cleanup or storage would be required. During operations, EEG electrodes would need to be in

contact with the scalp, usually in multiple locations. To conduct signals, a cream or gel typically must be used. The operator therefore must maintain a hairstyle that is conducive to electrode placement, and clean his or her hair after the shift. Further, the electrodes and the unit to relay data may restrict movement during operation and accidental shifts of the equipment may modify the signal or even cause electrodes to lose contact with the scalp. Eye tracking also has restrictions. Primarily, an operator must remain in view of the camera and not obstruct large portions of the face. These requirements may not be difficult to meet, as attention should be focused in the direction of the monitors for task performance. Further, loss of eye tracking is typically easy to detect and to correct. An operator will usually have to move back into the view of the camera or remove whatever is block the camera's view to regain tracking. EEG failures are likely to be more difficult to detect (artifact detection and removal) and time consuming to correct. Finally, automatic assessment of drowsiness using EEG is undependable (Schleicher, Galley, Briest, & Galley, 2008) and confounded by strong individual differences in EEG of alert operators (Kircher, Uddman, & Sandin, 2002) which presents major obstacles to the ready use of EEG for use in online systems. On the other hand, the use of eye tracking to detect fatigue online has already been successfully demonstrated (e.g., Kozak et al., 2005) and it has been identified as the most promising online gauge of fatigue (e.g., in driving; Wierwille et al., 1994). Taken together, eye tracking appears to be a more promising online gauge for present purposes.

Like EEG, eye tracking has been evaluated for online state detection in adaptive systems (e.g. workload; de Greef, Lafeber, van Oostendorp, & Lindenberg, 2009). A large body of research has linked eye tracking metrics to states of fatigue (R. M. Stern, Ray, & Quigley, 2001). Common metrics used to gauge fatigue relate to blink rate and eye lid closure duration (e.g., Brookings et al., 1996; Schleicher et al., 2008; J. A. Stern, Boyer, & Schroeder, 1994; Verwey & Zaidel, 2000). Blink frequency may be associated with the weakening of attention-driven inhibition of blinks, and

increases in blink duration might reflect the deactivation and slowing of physiological processes (Schleicher et al., 2008). Blink interval may not correlate strongly with alertness (Johns, Tucker, Chapman, Crowley, & Michael, 2007; Schleicher et al., 2008), however, eyelid closure has been found to be a reliable gauge of fatigue. Indeed Erwin (1976) found it to predict onset of sleep more effectively than plethysmography, respiration rate, heart rate variability, skin conductance, electromyography (EMG), and EEG (Wierwille et al., 1994). Dinges et al. (1989) also found eye tracking metrics to be more effective that EEG for detecting fatigue.

One eye tracking method for gauging fatigue is fixation duration. Eye movements consists of frequent, quick movements called saccades, interspersed with periods of steady gaze called fixation (Fischer & Weber, 1993; Palmer, 1999) (p22-23) Young and Sheena, 1975). During fixations, perception and cognitive activity occur (Nuthmann, Smith, Engbert, & Henderson, 2010), and extended fixations can indicate difficulty extracting information (Kramer & McCarley, 2003; Palmer, 1999). Specifically, as a person struggles to maintain focus and attention with fatigue, fixations lasting 150-900ms, which are associated with cognitive processing, decrease. Fixations longer than 900ms, indicative of staring, and less than 150ms, which may relate to low level unconscious control but not deep processing, increase (Schleicher et al., 2008). Mean fixation duration does not relate to fatigue (Saito, 1992). It is important to note that research that has successfully gauged fixation duration has used EOG equipment capable of 1000 Hz recording of eye movement. Schleicher et al (2008) point out that 500-1000 Hz is typically required to record saccades. However, other research has successfully registered fixation duration mean differences with other manipulations using a 60 Hz tracker. For example, Miyao, Hacisalihzade, Allen, and Stark (1989) found differences in fixation duration related to reading from different quality and size of screen interfaces (using 10 second samples). A 60 Hz tracker (as we intend to employ) may be sufficient for distinguishing fixation duration categories (express, cognitive, and overlong) which Schleicher et al. (2008) successfully used

to detect fatigue effects. A limitation of this metric is that fixation duration may also be sensitive to workload independently of fatigue (Poole & Ball, 2006).

Another, prominent metric linked to fatigue is percentage of eye closure (PERCLOS; Wierwille et al., 1994) which may be particularly useful for detecting fatigue (Dinges et al., 1998; Knipling & Wierwille, 1994; Wang et al., 2006). It is considered a standard drowsiness gauge by many researchers (Kircher et al., 2002). PERCLOS is the proportion of time that a person's eyes are more than 80% closed and, at its conception, served as the definitional measure of drowsiness (Wierwille et al., 1994). Other research has since used a more liberal 70% closed version of the metric (Dinges et al., 1998). In the seminal work on PERCLOS, Wierwille et al. (1994) stipulated that PERCLOS of 7.5% or less indicated that a person was "awake", the 7.5% to 15% range was questionable, and PERCLOS greater than 15% indicated the person was "drowsy". Some more recent studies have used greater than 8% PERCLOS to define drowsiness (e.g., Kozak et al., 2005). PERCLOS is reflective of slow eyelid closures more than blinks. Blinks less than 500 ms in duration are usually excluded from the computation (e.g., Kozak et al., 2005). Lid closures greater than 500 ms are usually defined as microsleeps (Schleicher et al., 2008).

While PERCLOS has been successful for predicting performance on vigilance tasks, more so even than subjective sleepiness ratings (Dinges et al., 1998), it does have some drawbacks. McKinley, McIntire, Schmidt, Repperger, and Caldwell (2011) found the metric to be not robust enough to detect reliable indications of degradations in performance due to fatigue. Other recent findings suggests that lid closure behavior is influenced by individual differences (e.g., Schleicher et al., 2008; Van Orden, Jung, & Makeig, 2000) which may prove to be a challenge to constructing an online gauge based on normative relationships unless individual differences can be factored into the model. A more fundamental concern is that certain types of fatigue states may not correspond to lid closure behavior (Galley & Schleicher, 2004). Lal and Craig (2001) reported a study (O'Hanlon & Kelley,

1977), which found that drivers with eyes open registered EEG indicators of sleep state. In this state, drivers are able to keep their lane, but cannot respond adequately to unexpected events. So called "driving without awareness" (DWA) or "highway hypnosis" (e.g., Wertheim, 1978), is a well-known, but still not fully understood phenomenon (Karrer, Briest, Vöhringer-Kuhnt, Baumgarten, & Schleicher, 2005). Anecdotally this state involves the competent performance of basic tasks without conscious engagement. DWA can be induced by bright points of fixation (Briest, Karrer, & Schleicher, 2006) and highly predicable environments (Wertheim, 1978), such as monotonous roads. Although some have questioned the relevance of eye closure, this state is purportedly marked by changes in gaze behavior (Galley & Schleicher, 2004; Wertheim, 1978). Karret et al. (2005) found that as frequency of DWA events increased, amplitude and duration of saccades decreased. A control station for UAV operation may be at risk of inducing DWA states, as it offer very little environmental variation and requires operators to monitor backlit displays. More recent research has indicated that eye lid closure behaviors, such as blink duration, may, in fact, be associated with DWA (Briest et al., 2006). Nonetheless, it may be important to supplement PERCLOS in the event that lid closure does prove to be an insensitive gauge of this fatigue state.

With the limitations of PERCLOS in mind, this study examined another eye tracking metric, recurrence quantification analysis (RQA) which may be useful for quantifying the amount and characteristics of stereotyped behavior, like the eye movements said to accompany DWA. RQA is a method of quantifying the amount of patterning and dynamic structure in a time series (Russell, 2014). The rationale for RQA stems from the dynamical systems conception that the behavior of any complex system, which is variable but generally stable, is a consequence of interactions between highly variable underlying components and processes (Kloos & Van Orden, 2010). According to Taken's embedding theorem (Takens, 1981), the influence of these many components and their interactions on the collective behavior of the system make it possible for information about the

underlying dynamics of a potentially multivariate system to be gleaned from a single scalar time series (M. A. Riley, Balasubramaniam, & Turvey, 1999).

As mentioned above, severe fatigue can induce a driving hypnosis or DWA state which is characterized by eye movement stereotypy (Briest et al., 2006). Thus, finding high determinism in a time series may indicate an early onset of fatigue. Theoretically, heightened determinism in behavioral variance, is indicative of rigid over-control, or high constraint, which can produce stable behavior in a predictable environment, but lacks the flexibility to allow the system to adjust to unpredictable environmental features (Kloos & Van Orden, 2010). This principle is in line with descriptions of DWA which maintain that drivers in such a state remain competent in highly predictable environments but fail to react adequately to environmental changes (Wertheim, 1978). RQA of muscle fatigue has found that while recurrent behavior remains relatively constant with fatigue, the amount of determinism (metric described below) increases (Ikegawa, Shinohara, Fukunaga, Zbilut, & Webber Jr, 2000; Morana, Ramdani, Perrey, & Varray, 2009; Webber Jr, Schmidt, & Walsh, 1995). Naschitz et al. (2002) found determinism in heart rate to be predictive of chronic fatigue syndrome. Applied to eye tracking, RQA might be used to register increased determinism in gaze patterns in relationship to fatigue.

Reliance on Automation

Levels of Automation

Levels of automation (LOA) refer to the tradeoff between automation control and operator control of a task or system. Levels of automation have been formally defined in supervisory control (Sheridan & Verplank, 1978) and since elaborated and adapted (e.g. Miller & Parasuraman, 2003; Parasuraman et al., 2000; Parasuraman & Wickens, 2008). Table 1 shows the ten levels of automation originally identified by Sheridan and Verplank (1978). These levels apply to context where the role of automation is to makes decisions, offers suggestions, and execute actions

(Parasuraman et al., 2000). These roles are consistent with automated decision aids utilized in UAV operation (c.f. Dzindolet et al., 2003; Eggers & Draper, 2006; Hanson & Harper, 2000) and will be the primary focus of the present effort.

Table 1

Levels of automation (LOA) of Sheridan & Verplank (1978) as summarized by Miller and Parasuraman (2003)

Level	*Description*
1.	Human does it all
2.	Computer offers alternatives
3.	Computer narrows down to a few
4.	Computer suggests a recommended alternative
5.	Computer executes alternative if human approves
6.	Computer executes alternative; human can veto
7.	Computer executes alternative and informs human
8.	Computer executes alternative and informs human only if asked
9.	Computer executes alternative and informs human only if it decides to
10.	Computer acts entirely autonomously

Note. Level 1 indicates lowest LOA, level 10 the highest LOA.

One type of automation that can use different levels of automation is decision aids. Decision aids aim to bolster situation awareness by enabling real time decision making without full and continuous human intervention, which would be untenable in multi-UAV operation (Hanson & Harper, 2000). Computers have the advantage of being able to combine data from multiple sensors into higher-level packets of information quickly. These packets range from perception of elements in the environment, to comprehending a situation, to inferring future events based on current sensor information (Cummings et al., 2007; Hanson & Harper, 2000). This aid might be applied to low-level decision making, such as target recognition, or high-level decision making, such as route

planning (Cummings et al., 2007; Drury & Scott, 2008). At an intermediate LOA, an operator can review the information and recommendations the system generates and enact or allow a course of action to be carried out (Cummings et al., 2007; Eggers & Draper, 2006). In decision aid research, levels of automation generally include full manual control, management-by-consent, management-by-exception, and full automation (Liu, Wasson, & Vincenzi, 2009; Ruff et al., 2002). Management-by-consent, typically involves an automated recommendation that must be confirmed or changed before any action is taken. Management-by-exception plans and executes actions automatically, unless the operator intervenes.

Although higher LOAs can reduce workload associated with managing the flight of multiple UAVs, it is crucial to keep the human operator in the loop of automated processes (Ruff et al., 2002). Computer algorithms are able to cope only with that which is programmed in anticipation of events that the automation might encounter and therefore are imperfect (Cummings et al., 2010; c.f. Silverman, 1992; P. J. Smith, McCoy, & Layton, 1997). When the human operator is removed entirely from the process, the repercussions can be catastrophic because humans can no longer correct failures in automated processes (Parasuraman & Riley, 1997). Such problems might occur when automation is used to prevent human operator error. A helpful illustration was provided by Parasuraman and Riley (1997): a system designed to slow an aircraft after landing might be programmed not to function unless the system senses that the landing gear is on the ground. The pilot is then prevented from inadvertently activating these systems in the air. However, the pilot is also prevented from intervening in the event that the landing gear sensors fail.

Another important implication of level of automation is its impact on situation awareness. Endsley (1996) points out that automation can improve situation awareness by reducing display complexity and improving the integration of information (Wiener, 1985) and by reducing high levels of workload (Billings, 1991) freeing the operator to process more information from more sources

more quickly (Curry & Ephrath, 1976). However the reduction of workload by automation is debatable (Endsley, 1996), and importantly, higher LOAs may result in a loss of situation awareness brought on by vigilance and complacency issues (Endsley, 1996; Endsley & Kiris, 1995; Miller & Parasuraman, 2007). Even when the operator is only removed from task implementation, performance recovery is much less successful in the event of an automation failure (Endsley & Kaber, 1999). This is especially risky in situations that involve human safety (Parasuraman & Wickens, 2008). Situation awareness decrements caused by this task disengagement can be mitigated with greater operator involvement (Parasuraman, Mouloua, & Molloy, 1996). Intermediate levels of automation allow for fewer out-of-the-loop problems while also improving situation awareness and lowering operator workload (Endsley & Kaber, 1999; Parasuraman & Wickens, 2008).

Previous research (Ruff et al., 2002) and an in house pilot study show that management-by-consent has performance advantages over higher and lower levels of autonomy in simulated UAV decision making tasks. For the scope of the present effort, it is necessary to control for LOA by holding it constant at a management-by-consent level for a high priority decision-making task in order to investigate the impact of fatigue, automation reliability, and trust on reliance. In a future effort, management-by-consent will serve as the default LOA from which online adjustments will be made in response to online assessments informed by the present effort.

Complacency, Reliability, Trust, and Reliance

The benefits of automation may be greatest in high-workload periods of multi-UAV operation were active fatigue is an issue (Hancock & Desmond, 2001). Here, automation can relieving the operator of some of the burden of multitasking by automatically carrying out tasking where human involvement is not crucial. This frees the operator to focus on tasks that require more human involvement and allows the operator to maintain better situation awareness. In addition to

reducing workload, automation can also increase operator situation awareness by providing more comprehensive packets of information to aid higher-level decision making.

Unfortunately, automation has the potential to exacerbate passive fatigue during low activity periods by reducing operator involvement in system tasking, relegating them to a passive monitoring role (Warm, Parasuraman, et al., 2008; see above for review). Fatigue may incline the operator to become complacent and allow the automation to do most of the work (Parasuraman et al., 1993) as an energy conservation strategy (Sauer et al., 2003). For example, at the LOA of "management-by-exception", where the machine chooses and carries out a response unless the operator intervenes, a fatigued operator may become complacent and reluctant to exert the effort necessary to check the utility of the automation continually. Further evidence for complacency is offered by Neubauer et al. (2011) who found that fatigued drivers tended to engage automation when it did not enhance performance.

Complacency has long been implicated in aviation accidents (Singh, Molloy, & Parasuraman, 1993) despite an early lack of consensus concerning its definition. One early and oft cited definition by Billings, Lauber, Funkhouser, Lyman, and Huff (1976) in the NASA Aviation Safety Reporting System (ASRS) defines complacency is a "self-satisfaction which may result in non-vigilance based on an unjustified assumption of satisfactory system state" (Parasuraman et al., 1993). Singh et al. (1993) identify reliance and trust as two components of complacent attitudes toward automation, but suggest that pilot attitudes toward automation are not enough to induce complacent behavior. Instead, contextual factors must also be at play for complacent behavior to occur, including fatigue due to poor sleep and long flights as well as fatigue due to high workload. These two sources of fatigue are consistent with context factors that induce passive and active fatigue respectively.

Although Singh et al. (1993) identifies reliance and trust as different components of complacency, others have focused on the impact of trust on reliance (Lee & Moray, 1992; Lee &

See, 2004; Parasuraman & Riley, 1997). Often this relationship is described in terms of trust's role as mediator between automation reliability and reliance on automation (Lee & Moray, 1992, 1994; Lee & See, 2004; Parasuraman & Wickens, 2008). Generally, as the reliability of the automation increases, so does trust in the automation, which encourages reliance. Further, trust may encourage complacency behaviors. Muir and Moray (Muir & Moray, 1996) found that higher levels of trust corresponded to reduced monitoring of the automation. Lee and See (2004) describe the relationship between trust and reliance in terms of calibration. Calibration refers to the balance between the amount of trust an operator puts into the automation and the true capability of the automation, or its trustworthiness. Overtrust occurs when trust in the automation exceeds the automation's actual capability, and distrust is when the amount of trust in the system is less than the system's capabilities. The influence of trust on reliance is moderated by its relationship with operators' perceived ability to perform tasks manually (Lee & Moray, 1992, 1994; Singh et al., 1993). Specifically, if operators' perception of their own ability exceeds their trust in the automation's performance, they will rely on automation less. Conversely, if operators' believe their ability to perform a task is less than their trust in the ability of the automation to perform the task, they will rely on automation more.

While it is apparent that fatigue impacts reliance, notably by inducing complacent behavior (see above), the mechanism by which this impact occurs has not been clearly identified. Indeed, models of reliance (Parasuraman & Mouloua, 1996) make no mention of a relationship between fatigue and trust in automation, an important driver of reliance. In the former case, there is evidence that reliability and other features of the automation, such as the level of variability in reliability (Parasuraman et al., 1993), etiquette (Parasuraman & Miller, 2004), emotion (Merritt, 2011), and knowledge of why automation failures occur (Dzindolet et al., 2003), influence trust in automation which drives reliance. However, there seems not to be any clear evidence that trust in automation is

directly influenced by fatigue. However, fatigue may impact trust in automation indirectly by reducing executive functioning. Maintaining focus for extended periods of time requires high levels of self-regulation, which can deplete executive functioning resources (Baumeister, Bratslavsky, Muraven, & Tice, 1998; Baumeister, Vohs, & Tice, 2007; Schmeichel, Vohs, & Baumeister, 2003; Tice, Baumeister, Shmueli, & Muraven, 2007; Vohs, Baumeister, & Ciarocco, 2005; Webb & Sheeran, 2003). Executive functioning is important for attention switching, which is crucial for multitasking (Gilbert & Burgess, 2008). If the impact on task performance and monitoring causes operators to miss critical automation performance signals, such as an inappropriate recommendation, trust in automation may be inappropriate.

Reliance literature also fails to identify a relationship between fatigue and self-confidence. However, execute function may serve as a mediator here as well. Maintaining focus on an unstimulating task requires resource-intensive self-regulation. A fatigued operator who recognizes low executive functioning resources may increase reliance on automation as a load shedding strategy (Schulte & Donath, 2011). An operator may also adjust reliance to reduce resource costs associated with checking the reliability of the automation. For example, an operator might reduce reliance when automation is known to be unreliable and the cognitive overhead to check continually is deemed too high (Parasuraman & Riley, 1997). Conversely, an operator may increase reliance when the automation is highly reliable (Bailey & Scerbo, 2007) and detecting a single error would require high levels of effort. Thus, fatigue may adjust reliance to the extent that operators must reduce strain on executive function. Importantly, fatigue may also impact reliance by distorting operators' perceptions of their ability to perform a task (Brown, 1997; Matthews & Desmond, 2002) which might result in inappropriate reliance. The impact of fatigue on executive function may reduce operators' awareness of performance impairments, especially in a multitask context, which might inform self-evaluation. Consequently, an operator may not rely enough on automation in an active

fatigue situation when increasing automation may improve performance, or may rely too heavily on automation in a passive fatigue situation when monitoring ability is poor and increased task-engagement may improve performance. Other factors may also impact reliance, such as risk and motivation, but these considerations were outside of the scope of this project.

In summary, fatigue may exacerbate the harmful effects of automation through its impact on operator reliance. Fatigue increases the need for self-regulation, depleting executive function resources (Kaplan & Berman, 2010). Executive functioning is important for calibrating trust in automation and perceptions of ability. The effects of executive function depletion, and efforts to preserve executive functioning resources, impact reliance on automation. As a result, fatigue is an important consideration in the effort to optimize reliance.

Measurement of Reliance

Measurement. This study aims to take a multifaceted approach to the measurement of trust and reliance. The measurement of trust, as a mediator of reliability and reliance, and sometimes considered a proxy for reliance, will be discussed along with measurement of reliance.

Subjective Measures. For measuring reliance, subjective measures exist for assessing complacency and trust in automation. Singh et al. (1993) point out that complacency behavior is a product of both complacency-potential and appropriate context (including fatigue). No scales yet exist for measuring complacency behavior. This may be due to difficult measuring complacent behavior subjectively. Feedback would require introspection about a behavior that may be at least partially unconsciously driven under fatiguing conditions (c.f. DWA; Briest et al., 2006). Further, complacent behavior is not flattering and may be especially vulnerable to response bias. However, Singh et al. (1993) have developed a multi-dimensional scale for assessing dispositional complacency-potential, the Complacency-Potential Rating Scale (CPRS). This scale accounts for attitudes that contribute to complacency behavior reflecting opinions about automation in general,

confidence in automation, opinions about reliance in automation, trust in automation relative to manual methods, and the safety of using automation. Generally, higher opinions of automation in each of these components indicate high potential for complacency. The present study will use the CPRS to understand the extent to which individual differences drive reliance across different levels of automation reliability and fatigue. For an online gauge to be successful, it is important to be able to calibrate performance and physiological indices to these differences.

Other subjective measures focus on the trust component of reliance. Trust scales can gauge disposition toward automation (e.g. Jian, Bisantz, & Drury, 2000; Madsen & Gregor, 2000) or attitude in response to an interaction with a particular automation (e.g. Bailey & Scerbo, 2007; Lee & Moray, 1992). The present study will use a custom measures adapted from Madsen and Gregor's Human-Computer Trust Scale (HCT) to measure dispositional trust in automation.

Behavioral Measures. Complacency became a construct of interest based on its performance consequences in aviation and other domains (Singh et al., 1993). It makes sense therefore to consider the performance impact of complacency in any comprehensive investigation of reliance. Performance measures provide an unfiltered but imperfect gauge of reliance. Typically, performance measures of reliance make a tacit assumption about the motivation behind certain behaviors. For example, Parasuraman et al. (1993) used ability to detect automation failure as a performance indicator of complacency. Specifically, errors in detecting automation failures in a system with a constant level of reliability was worse than detection when reliability varied. The assumption is that operators grew complacent when the reliability of the system was stable. The authors suggested that the results could be explained in terms of trust; variable automation may have provoked a more skeptical attitude toward the automation. It may also be possible that fatigue explains these results. The variable reliability condition may have provided a more engaging experience for participants

shielding them from the impact of fatigue. With this interpretation, Parasuraman et al.'s measure may not have just have gauged willingness to use the automation, but situation awareness as well.

Parasuraman et al. (1993) argued that boredom (i.e. fatigue) was not a factor in their findings: "Complacency is often linked to boredom arising from operator underload in highly automated systems… However, our finding that monitoring of automation failure is poor under multitask but not single-task conditions indicates that automation-induced complacency is not necessarily associated with low workload." (p. 17). It is noted in more recent literature however (Hockey, 1997) that under fatiguing conditions, compensatory behavior may be able to maintain performance on a primary task but not on subsidiary activities. In the multitask experiment, the automation supported system monitoring task may have served as a secondary task to the two manual tasks. This would explain the marked differences in monitoring performance without eliminating boredom as a potential confounding factor.

This effort took a different approach to the measurement of reliance on automation, which relied on a volitional behavior to gauge the level of reliance on automation rather than task performance. That is, instead of measuring detection of automation failures to infer reliance on automation, this study measures operators' use of automation more directly. Here, reliance was operationalized as the use of recommendations made by an automated decision aid (agreement). Agreement rate has been identified as a useful and even ideal measure of dependence on automation (Rice & Geels, 2010). Several studies have utilized agreement with automation to gauge reliance behavior (Madhavan, Wiegmann, & Lacson, 2006; Pak, Fink, Price, Bass, & Sturre, 2012; Rice & Geels, 2010; Rice, Keller, Trafimow, & Sandry, 2010; Wiegmann, 2002; Wiegmann, Rich, & Zhang, 2001). The general logic behind the use of agreement is that compliance drives agreement rates (Dixon & Wickens, 2006; Dixon et al., 2007; Rice et al., 2010). This method builds on the knowledge that operators may detect errors at a similar rate at all levels of automation reliability

(Parasuraman et al., 1993). Thus, an operator would detect more failures in a low reliability condition, which may affect willingness to rely on the recommendations made by the automation.

Psychophysiological Measures. The frequency and duration of visual inspections can be used to infer cognitive processing. For example, Galesic, Tourangeau, Couper, and Conrad (2008) found that participants' visual scanning reflected a bias in the amount of processing of response options when completing a survey. Specifically, the primacy effect (tendency to respond using the first options presented) corresponded to amount of participant eye gaze, which suggests that gaze frequency and duration can be used as a proxy for amount of processing.

Reliance in automation may also be expressed in eye gaze behavior by the amount of time a person fixates on tasks supported by automation. Wickens, Dixon, Goh, and Hammer (2005) found that participants dwelled on automated tasking areas more when an alerting automation had a tendency to miss critical signals than when the automation was reliable. Similarly, trust in an automated system can be registered as longer fixations on automated recommendations relative to other options. Guan and Catrell (2007) discovered that participants gazed almost exclusively at the top few results in a search engine task. Participants found the critical search result much more often when it was located closer to the top of the results list. Notably, this primacy effect was much more pronounced than that found in Galesic et al.'s (2008) survey investigation. These results suggest that participants trusted the automation to return valid options higher on the list of results. Despite these precedents, there has yet to be an investigation of trust or reliance in automated decision aids in a dynamic performance environment. Taken together, the dwell-based evidence reviewed above suggests that reliance on automation can be inferred from the amount of time participants spend inspecting the tasking area. Specifically, the less time participants spend reviewing tasking, the more reliant they are on automation.

Research on adaptive automation rarely investigates the use of eye gaze, but research on workload has successfully demonstrated the use of metrics such as pupil diameter (de Greef et al., 2009) and fixation dispersion (Fidopiastis et al., 2009) for triggering interventions. The use of eye gaze to detect trust and reliance in an automated decision aid is based on the idea that the extent to which a person scans an automated supported task area can be interpreted as an indirect indicator of trust in automation (Flemisch & Onken, 2000; Parasuraman et al., 1993). This effort will investigate the extent to which gaze duration might correspond to trust in and reliance on an automation decision aid.

Goal

The broad goal of this research was to support the development of adaptive interfaces that detect threats to optimal reliance. The findings were intended to support the online adjustment of the level of automation support to optimize reliance.

Aim 1: To evaluate candidate metrics for gauging of operator fatigue

The overarching goal of this project is to preempt suboptimal reliance and its undesirable performance consequences. Previous work has identified fatigue as a major cause of suboptimal reliance. Because there is usually a lag between the onset of fatigue and performance decrements, detection of fatigue may serve as an early warning signal for the onset of suboptimal reliance. To this end, the first aim is to determine which of the online metrics, i.e. fixation duration, PERCLOS, and percent determinism via RQA, could be validly used for gauging fatigue during UAV operation. The task characteristics of the experimental trial (monitoring for infrequent signals) as well as the duration of the trial (2 hours) were expected to induce increasing passive fatigue with time on task. It was hypothesized that each metric would show temporal trends indicative of fatigue during the 2-hour task run. Specifically, the percent of fixations per minute that were between 150ms and 900ms would decrease while <150ms and >900ms fixations increase, percent of time eyes were closed

(PERCLOS) would increase, and percent of recurrences that were part of a deterministic trend (RQA) would increase. It was also hypothesized that these measures of fatigue would be associated with performance and reliance metrics. Finally, as these were all purported measures of fatigue, it was hypothesized that they would all correlate with subjective fatigue measured post-task.

Aim 2: To assess the impact of automation reliability

The second aim was to understand how automation reliability influences subjective and objective responses. To this end, participants completed a mission aided by either a low reliability automated decision making aid (60% accurate) or a high reliability automated aid (86.7% accurate). It was hypothesized that subjective trust in automation (Human-Computer Trust Scale) would be lower for the low reliability condition. Further, it was hypothesized that operators would distrust the automation and consequently allocate more effort to monitoring the automation, or to doing the task without the assistance of the automation. Correspondingly, it was hypothesized that the amount of time participants fixate on the automated task area would be greater in the low reliability condition than in the high reliability condition, as a consequence of the participant allocating more attention to the task. Further, this extra effort was hypothesized to result in an increase in workload (NASA-TLX), higher task engagement (DSSQ), and lower objectively measured fatigue (fixation duration, PERCLOS, percent determinism via RQA).

Aim 3: To understand how impact of reliability changes with fatigue

The third aim was to understand how the impacts of automation reliability described in Aim 2 change with time on task, as fatigue develops. It was hypothesized that reliance would increase with time on task in the high reliability condition (complacency), but that reliance would decrease when the automation was unreliable (disuse). Further, it was hypothesized that fatigue would be less pronounced when the automation was less reliable; operators who felt compelled to check the automation or allocate effort to the tasks were likely to be more engaged in the task. Specifically,

those compelled by unreliability to attend more to the task would maintain a stable level of task engagement whereas those with a reliable automation would continue to grow less engaged. It was hypothesized, therefore, that the disparity in fatigue between low and high reliability (as gauged by eye tracking metrics of fixation duration, PERCLOS, percent determinism via RQA) would grow greater with time on task.

Aim 4: To build fatigue and complacency detection algorithms

The final aim was to build two algorithms to gauge both fatigue and suboptimal reliance based on the eye tracking outcomes. The first would be based only on PERCLOS and percentage of fixation duration (dwell time) that occurs within the automated target area. PERCLOS is a well understood metric and dwell time within the target task is practical to measure online. The second, more complex algorithm was to add eye tracking measures that may be more difficult to implement, fixation duration and percent determinism via RQA. Fixation duration may be difficult to implement because it traditionally requires much higher sampling rates than the eye tracker available for this project (60 Hz). The current method based on Schleicher et al. (2008) however, may not require such a high sampling rate. The use of percent determinism via RQA in this context is novel and untested for gauging fatigue online. It was hypothesized that both algorithms would predict subjective fatigue. In line with the goal of the study, it was also hypothesized that these algorithms would predict task performance on critical tasking. Finally, it was hypothesized that the complex algorithm would provide improved prediction over the simple algorithm, provided that it proves feasible to develop this more complex algorithm.

METHODS

Power Analysis

Based on calculations using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), a sample size of $N = 80$ was required to detect a medium effect size ($\eta^2_p = .09$) with a power level of 0.95 and an α of 0.05 for the mixed model ANOVAs used in the analysis. For the regressions, to test two models using four anticipated predictor variables in the complex model, the sample needed to be $N = 107$ for detection of a medium effect size ($f = 0.15$). Therefore the target sample size was $N = 110$. However, with the eye tracking quality issues experienced during data collection, this number was amended upward.

Participants

One hundred and thirty-one students from the University of Central Florida undergraduate psychology participant pool were recruited through the SONA Systems recruiting tool. Students received class credit for participation. The age range of participants was between 18 and 40 and no vulnerable populations were targeted. Populations vulnerable to stress manipulations were excluded, as were participants who were not able to perform required tasking (uncorrected vision impairment, lack of fluency in English, physical disability preventing competent use of the mouse and keyboard). A total of $N = 131$ (50 women, 81 men, $M_{age} = 19.86$, age range: 18-31 years) participants were used for all performance and subjective-based analysis. To be eligible for analysis, eye tracking data had to score at least a 2 (one eye tracking) or 3 (full tracking) 75% of the time on the faceLAB quality scale (range: 0-3). Although this 75% criterion is liberal, only eye tracking data for $N = 36$ (11 women, 25 men, $M_{age} = 19.64$, age range: 18-27 years) to $N = 39$ (13 women, 25 men, $M_{age} = 19.56$, age range: 18-27 years) were usable due to various technical issues, including poor edge tracking with two large monitors.

Apparatus

Lab Space

Participants were be seated at a desktop workstation (Figure 2). The simulator was run on a custom-built iBuyPower desktop powering two 1920 x 1200 pixel monitors and two speakers. A second HP desktop powered the eye tracker. This system supported the two monitors just mentioned as well as a smaller monitor at the researcher's station for monitoring participants' eye tracking quality during the experiment.

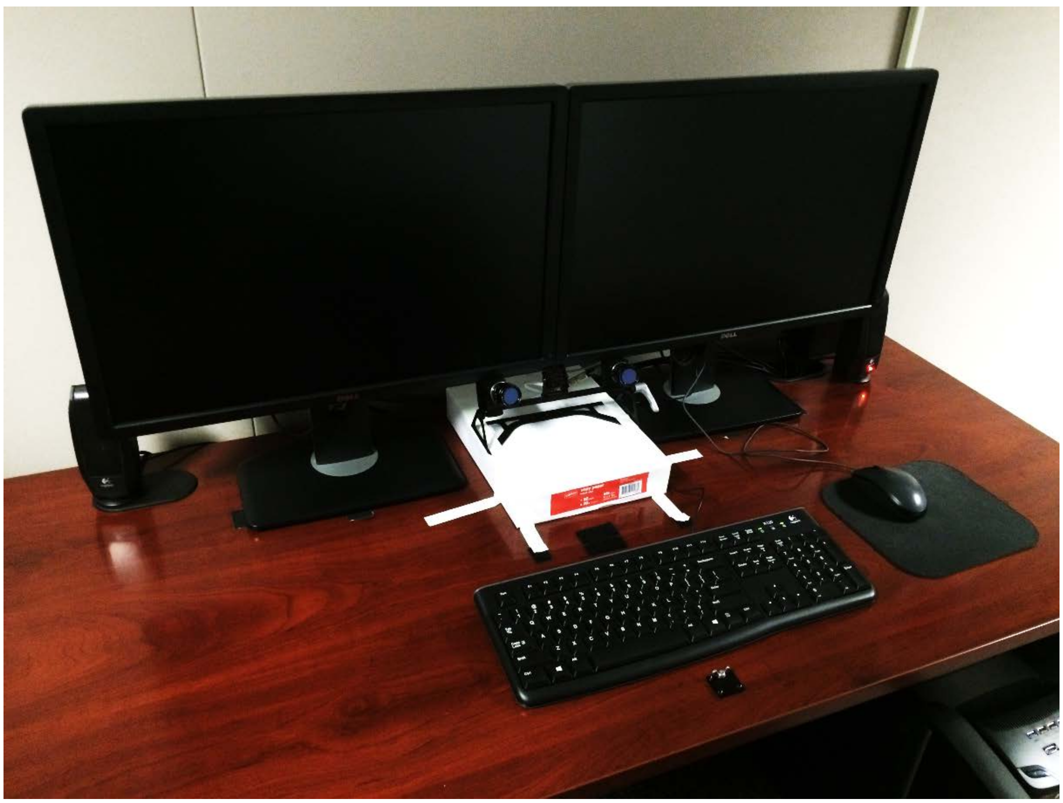


Figure 2. Participant desktop workstation.

Simulator

This effort called for a multiple UAV control station simulator that recreated a multitasking environment similar to that which an operator would experience. Further, it had to be possible to

change the event rate of tasking, set the duration of scenarios, and manipulate the level and reliability of automation for various tasks. The ALOA simulator, developed by OR Concepts Applied (ORCA; Johnson et al., 2007) met these requirements and has been employed in multiple studies of UAV automation (e.g. Calhoun et al., 2011; Kidwell, Calhoun, Ruff, & Parasuraman, 2012; Ruff & Calhoun, 2011) a previous experiment in my laboratory (Lin et al., 2015). ALOA supports several tasks, eight of which were used in the proposed effort (Table 2) to create a multitasking environment. The location of these tasks is shown in Figure 3.

Table 2

ALOA Simulator Tasking

	Task Type	Priority	Operator Action	Automation	Available Measures
1.	Allocation & Rerouting	None	None; Automated	Full Automation; No operator intervention permitted	N/A
2.	Image Analysis	1	Identify number of green diamond targets and click ID button for correct number	Recommendation only; No action taken without operator response	Response time and accuracy
3.	Weapon Release Authorization	1	Identify whether targets are correctly identified and click ID button to authorize strike if correctly identified	Recommendation only; No action taken without operator response	Response time and accuracy
4.	Unidentified Aircraft	3	Click red plane symbol when presented on map	None	Response time
5.	Digit Pairs	4	Determine whether two digits meet two criteria. Response is true or false	None	Response time and accuracy
6.	Audio Chatter	4	Identify color number combination prompted if call sign is named; ignore if not named	None	Response time and accuracy
7.	Health/Status	4	Click on yellow or red colored light	None	Response time
8.	Chat Questions	4	Answer question in chat window using information contained in vehicle status windows.	None	Response time and accuracy

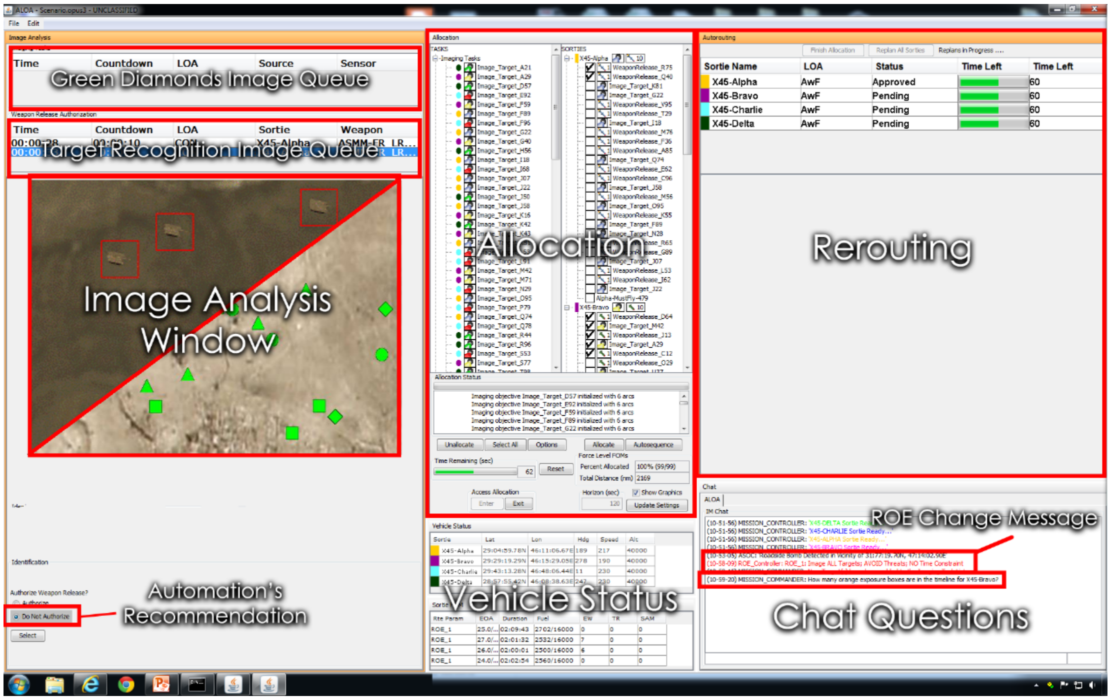
Note. This effort will not use performance outcomes for 4th priority tasks.

The frequency of tasks were low and allocation and routing tasking were completely automated to create a low workload situation. The experimental scenario lasted two hours to induce passive fatigue (based on pilot testing). An automated decision aid supported Image Analysis and Weapon Release Authorization tasking, which shared top priority status (see the left side of panel 'a'

in Figure 3). As these tasks became available, they stacked up as rows in two different queues at the top of the image analysis section. Participants had to open the tasks individually to complete them. When a participant clicked on a task in the queue, an image taken by an aircraft was displayed below the queue. The participant had to count green diamonds in the Image Analysis task or, for the Weapon Release Authorization task, ensure that a target recognition system had correctly identified target tanks. The Weapon Release Authorization images were degraded substantially on the basis that tasks with visually degraded stimuli tend to show rapid vigilance decrement (Warm, Dember, & Hancock, 1996).

For these tasks, the automation highlighted and selected a response recommendation. This recommendation was correct 60 percent of the time in the low reliability condition and 87.5 percent of the time in the high reliability scenario (reliability levels were comparable to those in the low and high reliability conditions of Parasuraman et al., 1993). The participant had to confirm or change the response to complete the task. A lack of response was considered an incorrect response.

a)



b)

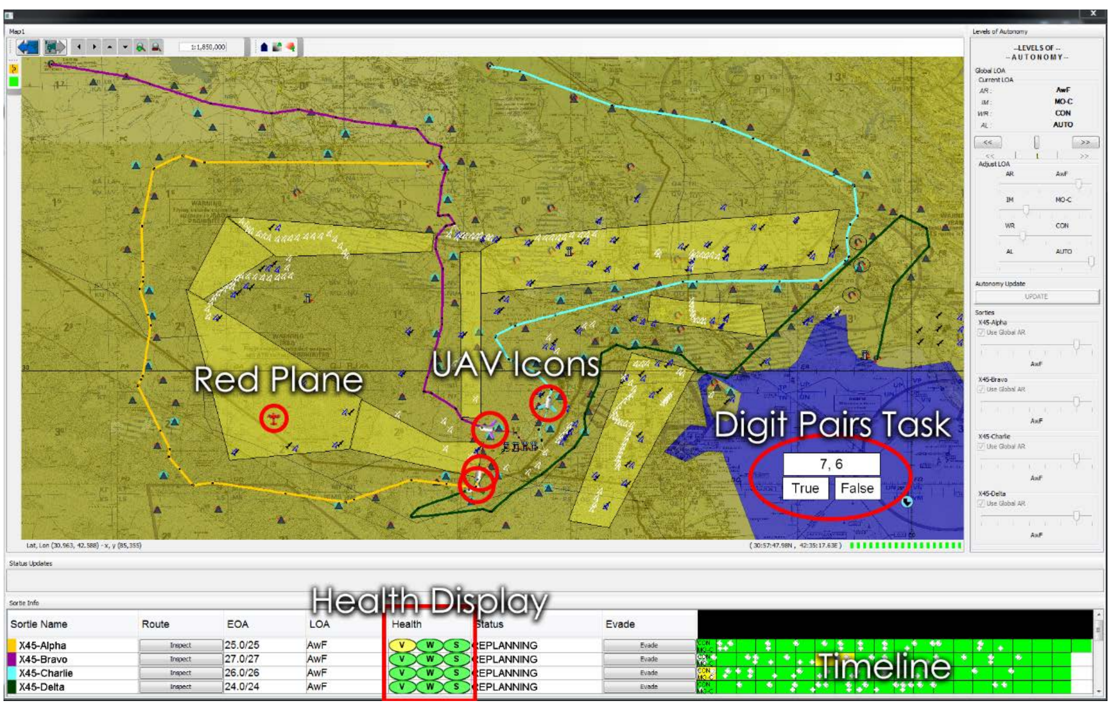


Figure 3. ALOA Simulator left (a) and right (b) panels.

Eye tracker

A faceLAB 5 desk-mounted eye tracking system by Seeing Machines was used to capture ocular data for eye tracking metrics (Figure 4). This system captured gaze direction, eye closure, facial gestures and head position. It also featured a built in PERCLOS assessment that used absolute eyelid position (as opposed to bright pupil or corneal occlusion approaches). Data was captured at 60Hz using two stereoscopic infrared filtered cameras and an infrared light source located between them. A quality monitor grades data to aid with data screening, which were especially useful with the dual monitor arrangement.

Figure 4. faceLAB 5 desktop eye tracker by Seeing Labs.

Subjective Measures

Demographics Questionnaire

The demographics questionnaire consisted of 20-items. It asked about a range of biographical information including, health, education, computer use, and video gaming experience (APPENDIX A: DEMOGRAPHICS QUESTIONNAIRE).

NASA-TLX

The NASA-Task Load Index (NASA-TLX; Hart & Staveland, 1988) is a multi-dimensional subjective assessment of workload (APPENDIX B: NASA TASK-LOAD INDEX) which has been widely used in the UAV and automation research domain (e.g. Endsley & Kaber, 1999; Fidopiastis et al., 2009). It combines six workload-related factors: performance, mental demand, physical demand, temporal demand, effort, and frustration level. All factors were rated on a 0-100 visual analog scale. Performance is anchored by "good" and "poor" and all other factors are anchored by "low" and "high". Test-retest reliability is $r = .83$.

Dundee Stress State Questionnaire (DSSQ; short version)

The short version of the Dundee Stress State Questionnaire (DSSQ: Matthews et al., 2002) was administered to gauge symptoms of active and passive fatigue (

APPENDIX C: DSSQ-3 STATE QUESTIONNAIRE). The short DSSQ has been verified as a valid alternative to the full DSSQ (Helton, 2004). Participants completed a pre-task DSSQ to serve as a baseline prior to training and trial completion. They later completed a post-task DSSQ according to their state in the final 10 minutes of the main trial. The DSSQ assesses three higher order dimensions of subjective state in a task-performance context: task engagement, distress, and worry. Items are sampled from three primary dimensions of subjective experience, affect, motivation, and cognition. Participants responded to statements describing current emotional state

(e.g. *I was motivated to try hard at the task*.) using a 5-point Likert scale anchored by 0 = "Definitely false" and 4 = "Definitely true".

Human-Computer Trust Scale

The scale used in the present effort is based on the Human-Computer Trust Scale (HCT; Madsen & Gregor, 2000), which is a reliable (Cronbach's α = 0.94) measure of affective and cognitive components of trust in automation (APPENDIX D: HUMAN-COMPUTER TRUST SCALE). Items gauge confidence in automated systems and willingness to act on recommendations made by the automation through five constructs believed to impact level of trust in an automated decision aid: perceived reliability (R), perceived technical competence (T), perceived understandability (U), faith (F), and personal attachment (P). The version used in the proposed effort uses items from each (R3, F3, T3, U2, R4, F1, P4, & U3, in that order). It also included a direct item about level of trust in automation, "Overall, I trust the automation." Participants responded to items based on their experience with an automated system by indicating the extent to which they agreed with statements about their trust using a 5-point Likert scale anchored by 0 = "Extremely disagree" and 4 = "Extremely agree" (e.g. *I can rely on the system to function properly*).

Procedure

The investigator greeted participants in the lobby of Partnership II and lead them to the 306D lab space. There the researcher provided an informed consent to review (approximately 5 minutes). After verbal agreement to proceed with the study, participants were directed to turn off their cellphones and to remove and store any visible timepieces. Before any experimental activities, the eye tracker was calibrated to model their full head and eyes (approximately 10 minutes). Participants then completed the Demographics Questionnaire, the Complacency Potential Rating Scale, the 40 Mini-Marker Personality Scale, and a pretask short DSSQ to establish baseline level of stress and fatigue (approximately 20 minutes).

Training followed (approximately 60 minutes) consisting of a PowerPoint introduction to the ALOA interface and the tasks contained in each section of the simulated UAV control station, followed by a detailed explanation of steps to complete each task on a live training simulation. Once oriented, participants performed a practice trial that contained all tasking simultaneously while the researcher monitored and provided assistance as needed. A second practice was available to be run if the participant did not performing adequately, but this option was never required. When the practice was completed, the computer was restarted (approximately 5 minutes). During this time the participant was be allowed to use the restroom (leaving portable electronics in the lab space), stretch, and review a "cheat sheet" that has information on how to do each task and the priority with which they should be done. The cheat sheet was available during the full trial.

Once the computer was restarted the eye tracking was rechecked and started, and the participant completed the two-hour long trial. During this time, the participant was not permitted to interact with the researcher in any way unless there was an emergency (e.g. urgent need to use the restroom). The researcher monitored the eye tracker to ensure that the participant remained in frame. After the trial, the participant completed questionnaires relating to their trust in the ALOA automation (Human-Computer Trust Scale and Metrics for Trust in Automation), a workload assessment (NASA-TLX), and a short version of the post-task DSSQ (approximately 15 minutes). Participants were instruction to respond according to the last 10 minutes of the full trial for the DSSQ. Finally, the researcher debriefed the participant, answered any questions about the session, provided an optional psychology research survey, and dismissed the participant (approximately 5 minutes). Credit was granted according to the amount of time the participant participated. A minimum of 4 hours credit was awarded to all completed sessions.

RESULTS

Subjective Outcomes

Stress State

A 2 (pre vs. post) x 2 (low vs. high reliability) mixed model ANOVA was run for each stress state factor to determine whether the mission elicited a stress state pattern consistent with that of fatigue. Results revealed no significant changes in distress in either reliability condition (Figure 5). The mission did significantly reduce task engagement, $F(1, 129) = 197.64$, $p < .001$, $\eta^2_p = .605$. Finally, there was a significant interaction effect for worry, $F(1, 129) = 5.27$, $p = .023$, $\eta^2_p = .039$. Post hoc contrasts showed that there was a significant increase in the low reliability condition ($p <$ 001), but no change in the high reliability condition ($p = .570$).

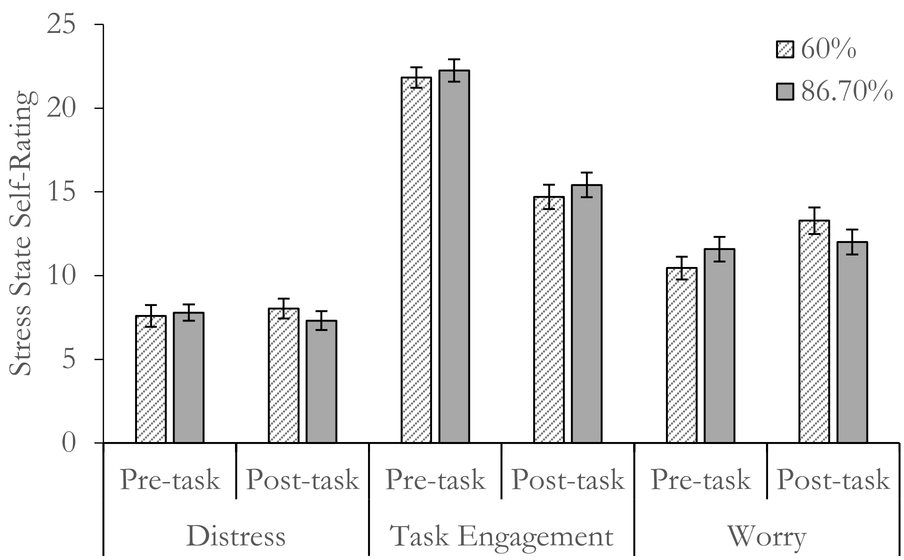


Figure 5. Self-reported stress state. Bars represent standard error.

Workload

To determine the nature of workload and the relative workload of each automation reliability condition, a 2 (low vs. high reliability) x 6 (workload factor) mixed model ANOVA was conducted. Mauchly's test indicated a violation of sphericity, $\chi^2(14) = 68.96$, p < .001. Degrees of freedom were corrected using Box's estimates of sphericity ($\varepsilon = .82$). The mission impacted some types of workload more than others (see Figure 6), $F(4.08, 526.69) = 31.15$, $p < .001$, $\eta^2_p = .195$. Specifically, Bonferroni corrected t-test comparisons indicated that perceived mental demand was significantly higher than all other types of perceived workload ($p < .001$). Conversely, perceived physical demand was the lowest of all types ($p < .001$). No other differences were found between workload types.

There was a small but significant main effect for reliability, $F(1, 129) = 6.38$, $p = .013$, $\eta^2_p = .047$. The low reliability condition elicited higher workload overall ($M = 39.65$, $SD = 16.64$) than did the high reliability condition ($M = 32.26$, $SD = 16.82$). There was no interaction between type of workload and reliability condition, $F(4.08, 526.69) = 0.50$, $p = .739$, $\eta^2_p = .004$.

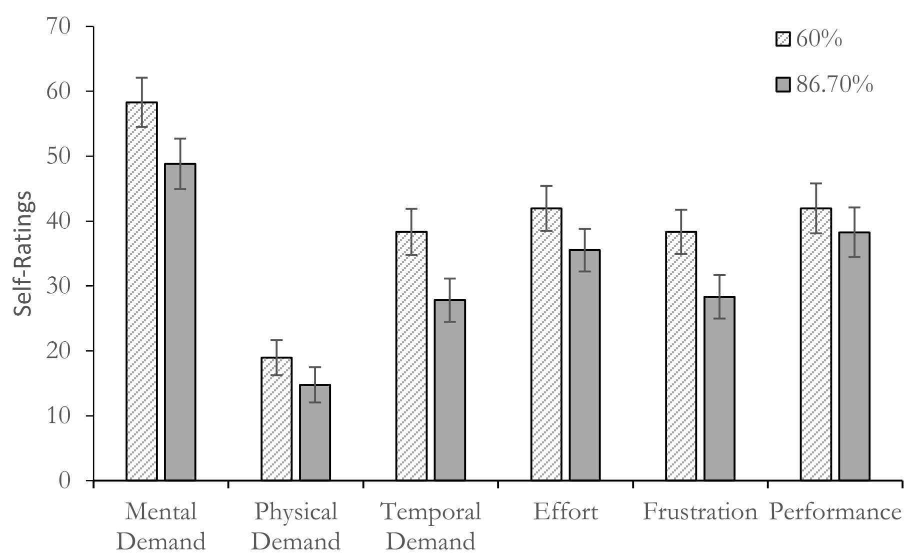


Figure 6. Workload factor ratings. Bars represent standard error.

Trust

Analysis of the Human-Computer Trust Scale trust revealed that participants trusted the high reliability automated aid (M = 21.67, SD = 6.05) more than the low reliability automated aid (M = 19.17, SD = 5.36), $t(129) = -2.50$, $p = .014$, $d = -0.44$.

Performance-Based Outcomes

Performance data for the Image Task and Weapon Release Authorization Task were used to understand the impact of time on task and reliability of automation on ability to perform the tasks and reliance on automation. Specifically, an 8 (time on task in 15 minute intervals) x 2 (low vs. high reliability) ANOVA was run for accuracy of responses and reliance on automation. Accuracy reflected the percent of total responses that were correct and reliance reflected the percent of total responses for which the participant agreed with the automation's recommendation, regardless of whether or not that recommendation was correct.

Accuracy

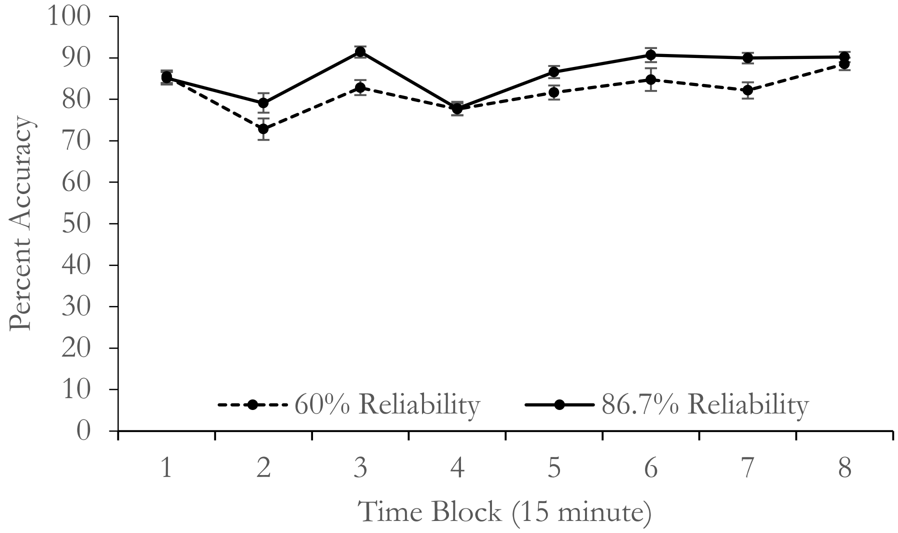
Image Analysis. Mauchly's test indicated a violation of sphericity, $\chi^2(27) = 103.01$, $p < .001$. Degrees of freedom were corrected using Box's estimates of sphericity ($\varepsilon = .80$). Participants performed better with a more reliable automated aid (Figure 7), $F(1, 129) = 9.52$, $p = .002$, $\eta^2_p = .069$. There was also a significant main effect for time, $F(5.57, 718.39) = 18.70$, $p < .001$, $\eta^2_p = .127$. Planned analyses revealed a significant linear trend, $F(1, 129) = 32.96$, $p < .001$, $\eta^2_p = .204$, indicating improvement in accuracy with time, as well as a significant quadratic trend, $F(1, 129) = 16.15$, $p < .001$, $\eta^2_p = .111$. While accuracy counting the green diamonds generally improved with time, it actually decreased slightly at first before accelerating toward the end of the session. Finally, results revealed a significant but weak interaction between reliability and time, $F(5.57, 718.39) = 2.41$, $p = .030$, $\eta^2_p = .018$.

The same analysis was performed for a subset of the data, which had usable corresponding eye tracking data so that these analyses could be compared with eye tracking analyses. Mauchly's test indicated a violation of sphericity, $\chi^2(27) = 42.59$, $p = .030$. Degrees of freedom were corrected using Box's estimates of sphericity ($\varepsilon = .71$). Again, participants performed better with a more reliable automated aid (Figure 8), $F(1, 37) = 4.22$, $p = .047$, $\eta^2_p = .047$. As before, there was also significant main effect for time, $F(5.00, 184.91) = 5.41$, $p < .001$, $\eta^2_p = .128$. Planned analyses revealed a significant linear trend, $F(1, 37) = 8.80$, $p = .005$, $\eta^2_p = .192$, indicating improvement in accuracy with time, and a marginal if not significant quadratic trend, $F(1, 37) = 2.97$, $p < .093$, $\eta^2_p = .074$. While accuracy counting the green diamonds generally improved with time, it actually decreased slightly at first before accelerating toward the end of the session. For this reduced sample, there was no significant interaction between time and reliability factors.

Weapon Release Authorization. As with the Imagining Task, Mauchly's test found a violation of sphericity, $\chi^2(27) = 126.12$, $p < .001$. Degrees of freedom were corrected using Box's estimates of sphericity ($\varepsilon = .74$). As with the Imaging Task participants performed better on the Weapon Release task with the more reliable automated aid (Figure 7), $F(1, 129) = 11.05$, $p = .001$, $\eta^2_p = .079$. Time again showed a main effect, $F(5.15, 663.75) = 27.52$, $p < .001$, $\eta^2_p = .176$, which in this instance seems due to a decline in performance with time, except for a recovery near the end of the mission. Planned analyses revealed a significant linear trend, $F(1, 129) = 13.42$, $p < .001$, $\eta^2_p = .094$, as well as a strong quadratic trend, $F(1, 129) = 26.77$, $p < .001$, $\eta^2_p = .172$. Accuracy declines before leveling off, then improving toward the end of the session. Finally, results revealed a moderate interaction between reliability and time, $F(5.15, 663.75) = 17.51$, $p < .001$, $\eta^2_p = .120$. A visual inspection of the means suggests that performance appears to differ more with time between the two reliability conditions.

For reasons outlined above, the analysis was repeated using a subsample of the data. Mauchly's test found a violation of sphericity, $\chi^2(27) = 49.22$, $p = .006$. Degrees of freedom were corrected using Box's estimates of sphericity ($\varepsilon = .67$). Unlike the full data sample analysis, there was no main effect for reliability condition. However, time did show a main effect similar to that of the full sample, $F(4.72, 174.57) = 7.93$, $p < .001$, $\eta^2_p = .177$, driven by significant linear, $F(1, 37) = 6.06$, $p = .019$, $\eta^2_p = .141$, and quadratic, $F(1, 37) = 7.89$, $p = .008$, $\eta^2_p = .176$, trends. Similar to the full sample, accuracy declines then levels off, but improves at the end of the mission. These findings are qualified by a significant interaction between reliability and time, $F(4.72, 174.57) = 5.74$, $p = .022$, $\eta^2_p = .134$. A visual inspection (Figure 8) of the means suggests that performance appears to differ more with time between the two reliability conditions.

a)



b)

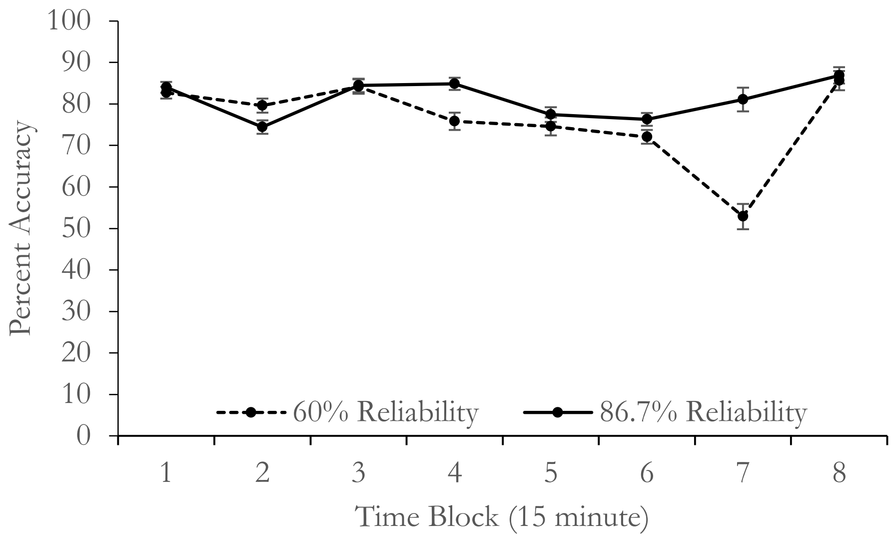


Figure 7. Accuracy for Image Analysis (a) and Weapon Release Authorization (b) tasks. Bars represent standard error.

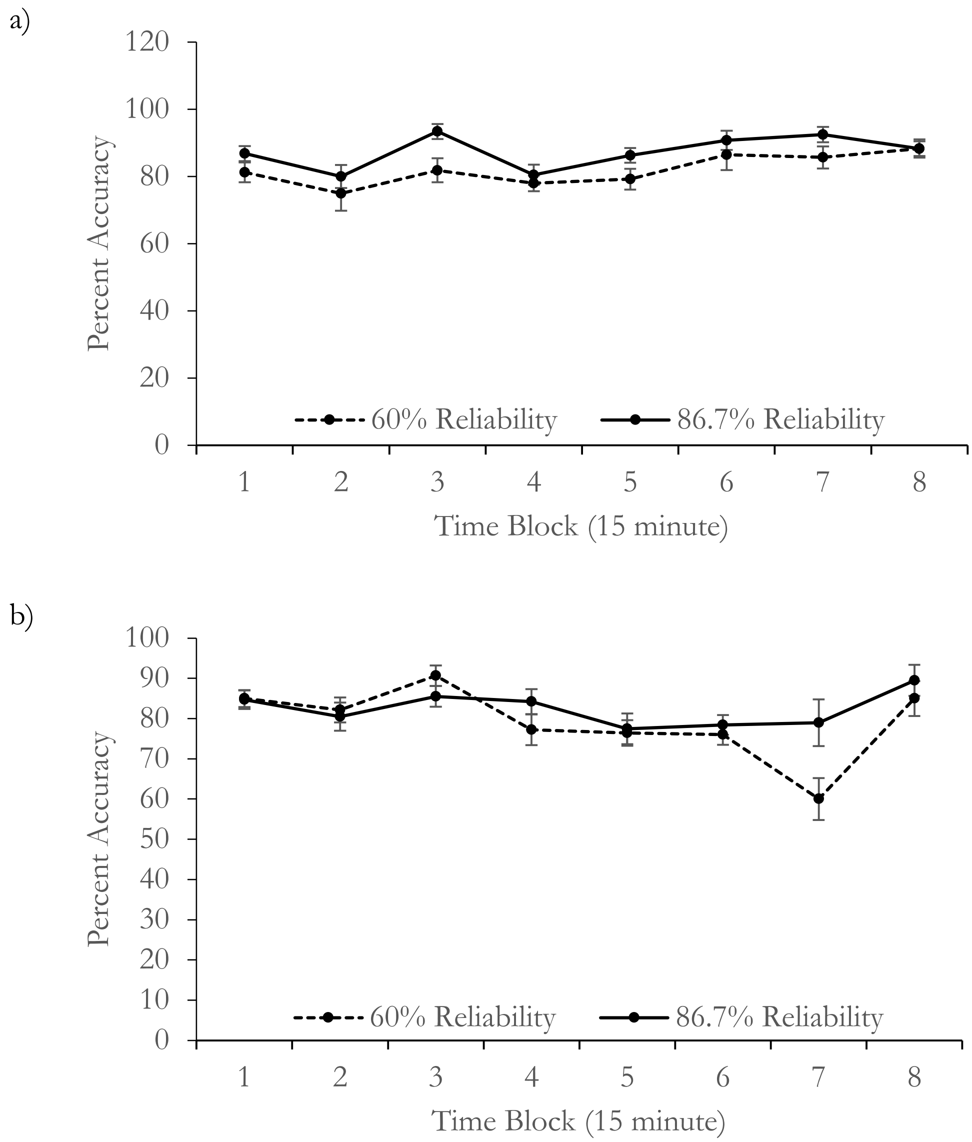


Figure 8. Subsample (N = 39) accuracy for Image Analysis (a) and Weapon Release Authorization (b) tasks. Bars represent standard error.

Reliance

Image Analysis. Mauchly's test indicated a violation of sphericity, $\chi^2(27) = 129.56$, $p < .001$. Degrees of freedom were corrected using Box's estimates of sphericity ($\varepsilon = .78$). Reliance behavior differed substantially based on the level of automation reliability (Figure 9), $F(1, 129) = 393.35$, $p < .001$, $\eta^2_p = .753$. Results also revealed a significant main effect for time, $F(5.44, 701.92) = 40.48$, $p <$

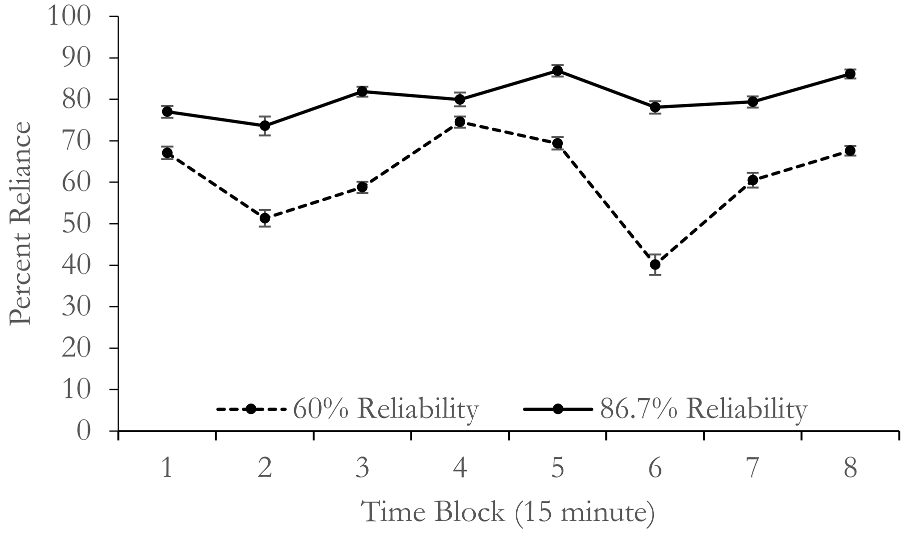
.001, $\eta^2_p = .239$. Planned analyses showed a slight linear trend, $F(1, 129) = 7.61$, $p = .007$, $\eta^2_p = .056$, but no quadratic trend. Both main effects were qualified by an interaction between reliability and time, $F(5.44, 701.92) = 19.41$, $p < .001$, $\eta^2_p = .131$. Reliance on the high reliability automation was more consistent than reliance on the low reliability automation.

As with the accuracy analyses, reliance analyses were run again using the subsample of participants who also had usable eye tracking data to facilitate direct comparison with eye tracking outcomes. Mauchly's test indicated a violation of sphericity, $\chi^2(27) = 55.61$, $p = .001$. Degrees of freedom were corrected using Box's estimates of sphericity ($\varepsilon = .69$). As before, there were strong main effects for reliability, $F(1, 37) = 192.82$, $p < .001$, $\eta^2_p = .995$, and time, $F(4.85, 179.50) = 17.38$, $p < .001$, $\eta^2_p = .320$. In contrast with the full sample, there was no significant linear or quadratic trend to explain the main effect for time. Main effects were still qualified by an interaction between reliability and time, $F(4.85, 179.50) = 8.24$, $p < .001$, $\eta^2_p = .182$. Although a pattern of interaction is not easy to discern visually, it was notable that reliance in the low reliability automation was more varied than reliance on the reliable automation (Figure 10).

Weapon Release Authorization. Again, Mauchly's test found a violation of sphericity, $\chi^2(27) = 126.35$, $p < .001$. Box's estimates of sphericity ($\varepsilon = .74$) were used to correct degrees of freedom. As with the imaging task, participants with the higher reliability automated aid were more reliant than those with the low reliability automation (Figure 9), $F(1, 129) = 205.66.$, $p < .001$, $\eta^2_p = .615$. A main effect for time revealed a steady decrease in reliance over time for both conditions, $F(5.14, 663.40) = 41.99$, $p < .001$, $\eta^2_p = .246$. Planned trend analysis indicated that this effect was linear, $F(1, 129) = 147.07$, $p < .001$, $\eta^2_p = .533$. Finally, these main effects were qualified by a significant interaction effect, $F(5.14, 663.40) = 12.46$, $p < .001$, $\eta^2_p = .088$. Reliance on low reliability automation was more erratic and saw a more pronounced decline than did reliance on high reliability automation.

For the analysis with the subsample with matching eye tracking data, Mauchly's test also found a violation of sphericity, $\chi^2(27) = 62.95$, $p < .001$. Box's estimates of sphericity ($\varepsilon = .64$) were used to correct degrees of freedom. Like the full sample, there were strong main effects for reliability, $F(1, 37) = 116.10.$, $p < .001$, $\eta^2_p = .758$, and time, $F(4.48, 165.64) = 18.67$, $p < .001$, $\eta^2_p = .335$. The time effect was supported by a significant linear trend, $F(1, 37) = 54.73.$, $p < .001$, $\eta^2_p = .597$. Participants relied less on the unreliable automation and reliance declined with time (Figure 10). Finally, these main effects were qualified by a significant interaction effect, $F(4.48, 165.64) = 7.26$, $p < .001$, $\eta^2_p = .164$. Reliance on low reliability automation was more varied and declined more quickly with time on task.

a)



b)

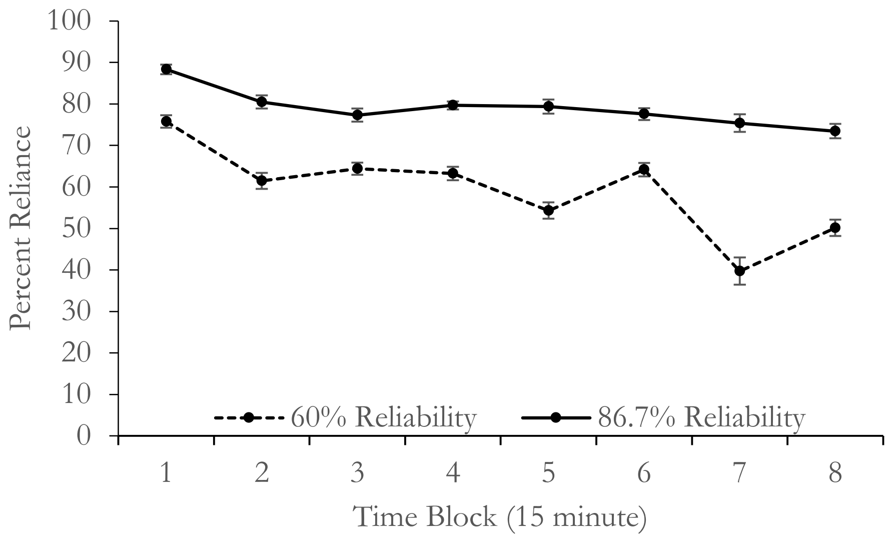


Figure 9. Reliance on automation in (a) Image Analysis and (b) Weapon Release Authorization tasks. Bars represent standard error.

a)

b)

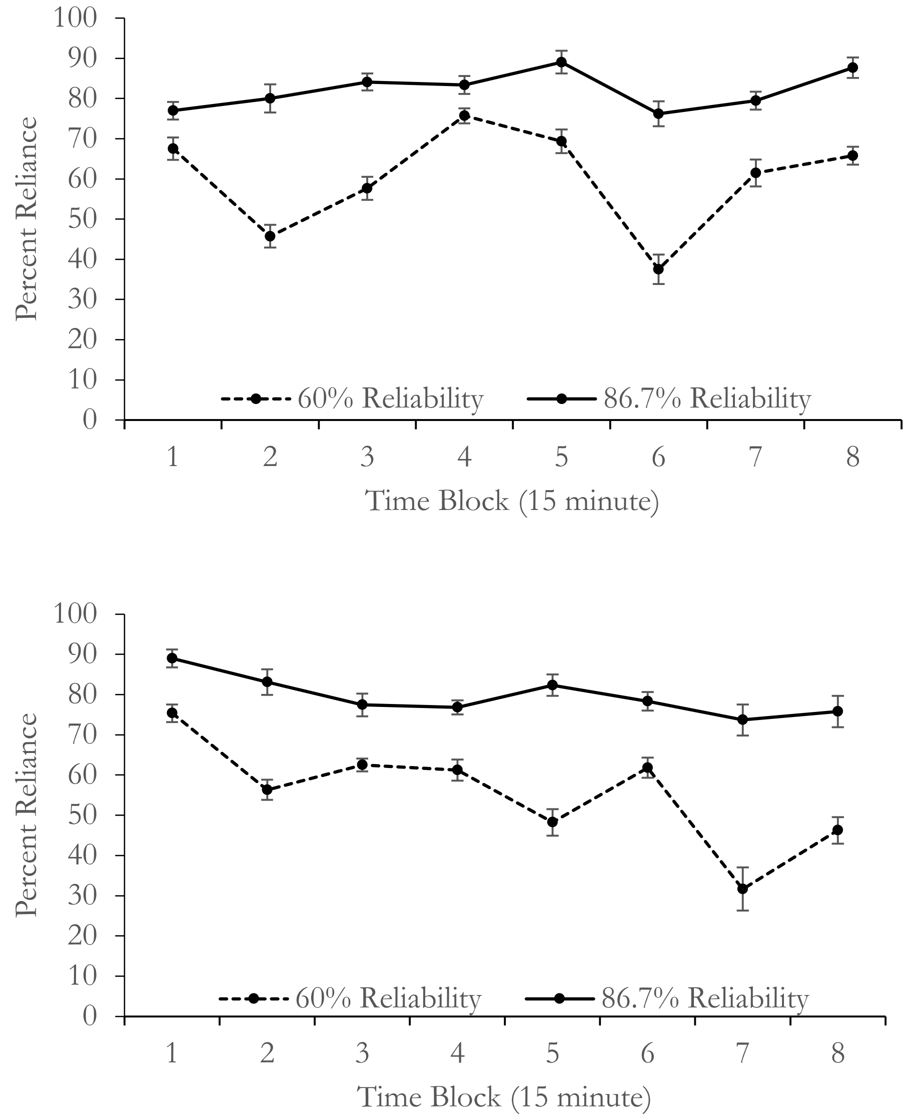


Figure 10. Subsample (N = 39) reliance for Image Analysis (a) and Weapon Release Authorization (b) tasks. Bars represent standard error.

Accuracy vs. Reliance

Finally, to understand the relationship between accuracy and reliance at each level of automation reliability, correlations were computed (Table 3). For both the Image Analysis and Weapon Release Authorization tasks, there was a negative relationship between accuracy and reliance in the low reliability condition particularly at the beginning and end of the mission. There

was a positive relationship between reliance and accuracy in the high reliability condition for both tasks. Notably, in the high reliability condition, the relationship abruptly reverses for Image Analysis task period six and Weapon Release Authorization task period seven.

Table 3
Correlations (r) between Accuracy and Reliance

	Period							
	1	2	3	4	5	6	7	8
60% Reliability								
Image Analysis	-.33**	-.38**	-.02	.08	-.24	-.93**	-.34**	-.15
Weapon Release	.02	-.36**	-.19	-.05	.19	-.12	-.97**	-.79**
86.7% Reliability								
Image Analysis	.68**	.36**	.60**	.59**	.13	-.38**	.48**	.22
Weapon Release	.56**	.26*	.65**	.36**	.40**	.55**	-.54**	.36**

* $p < .05$, ** $p < .01$

Time and Automation Reliability Analyses for Candidate Eye Tracking Indices of Fatigue

Eye tracking quality was problematic. Only 39 of the 131 participants had sufficient eye tracking quality for analysis. To be eligible for analysis, eye tracking data had to score at least a 2 (one eye tracking) or 3 (full tracking) 75% of the time on the faceLAB quality scale (range: 0-3). Unfortunately, even data that met this criterion contained only about half of the number of fixation events typically found in past research using the same eye tracking system. This is likely due to the fact that two 24 inch monitors were used and the participant was required to sit at a desk near those monitors for interaction. Calibration records often showed that the eye tracker was unable to provide good quality tracking near the far left and right edges of the two screen setup. The tops of the edges were particularly problematic. Therefore, the following eye tracking-based outcomes must be interpreted cautiously with these shortcomings in mind.

To determine whether or not eye tracking measures registered changes with time on task and reliability of automation, 8 (time on task in 15 minute intervals) x 2 (low vs. high reliability)

ANOVAs were run for each metric. Planned follow-up analysis for time on task included linear and quadratic trend analysis.

PERCLOS

PERCLOS data were analyzed using 20% Winsorized means for each period. Blinks as were identified and removed from analysis using faceLAB's built-in blink identification algorithm, which employs a maximum blink duration of 350 ms. Thus only eye closures longer than 350 ms were taken into account. Two sets of PERCLOS analyses were run, one using eye closure of at least 80%, and another using eye closure of at least 70%.

PERCLOS using 80% eye closure. Mauchly's test indicated a violation of sphericity, $\chi^2(27) = 175.71$, $p < .001$. Degrees of freedom were corrected using Box's estimates of sphericity ($\varepsilon = .50$). Results revealed a main effect for time on task, $F(3.52, 130.12) = 5.67$, $p = .001$, $\eta^2_p = .133$, which follow-up trend analysis indicated was mostly linear, $F(1, 37) = 9.89$, $p = .003$, $\eta^2_p = .211$ but somewhat quadratic, $F(1, 37) = 5.18$, $p = .029$, $\eta^2_p = .123$. Percentage of eye closure increased near the beginning of the mission, and abruptly began to decline steadily from time block 4 onwards (see Figure 11). There was no interaction between time and reliability.

PERCLOS using 70% eye closure. Mauchly's test indicated a violation of sphericity, $\chi^2(27) = 196.45$, $p < .001$. Degrees of freedom were corrected using Box's estimates of sphericity ($\varepsilon = .47$). As with PERCLOS using 80% eye closure, PERCLOS with 70% closure showed no reliability effect, but a main effect of time on task, $F(3.27, 121.10) = 4.91$, $p = .002$, $\eta^2_p = .117$, driven by linear, $F(1, 37) = 6.58$, $p = .015$, $\eta^2_p = .151$, and quadratic, $F(1, 37) = 4.45$, $p = .042$, $\eta^2_p = .107$, trends. The pattern for 70% closure was very similar to the 80% closure pattern, except that the linear trend was less pronounced (see Figure 11).

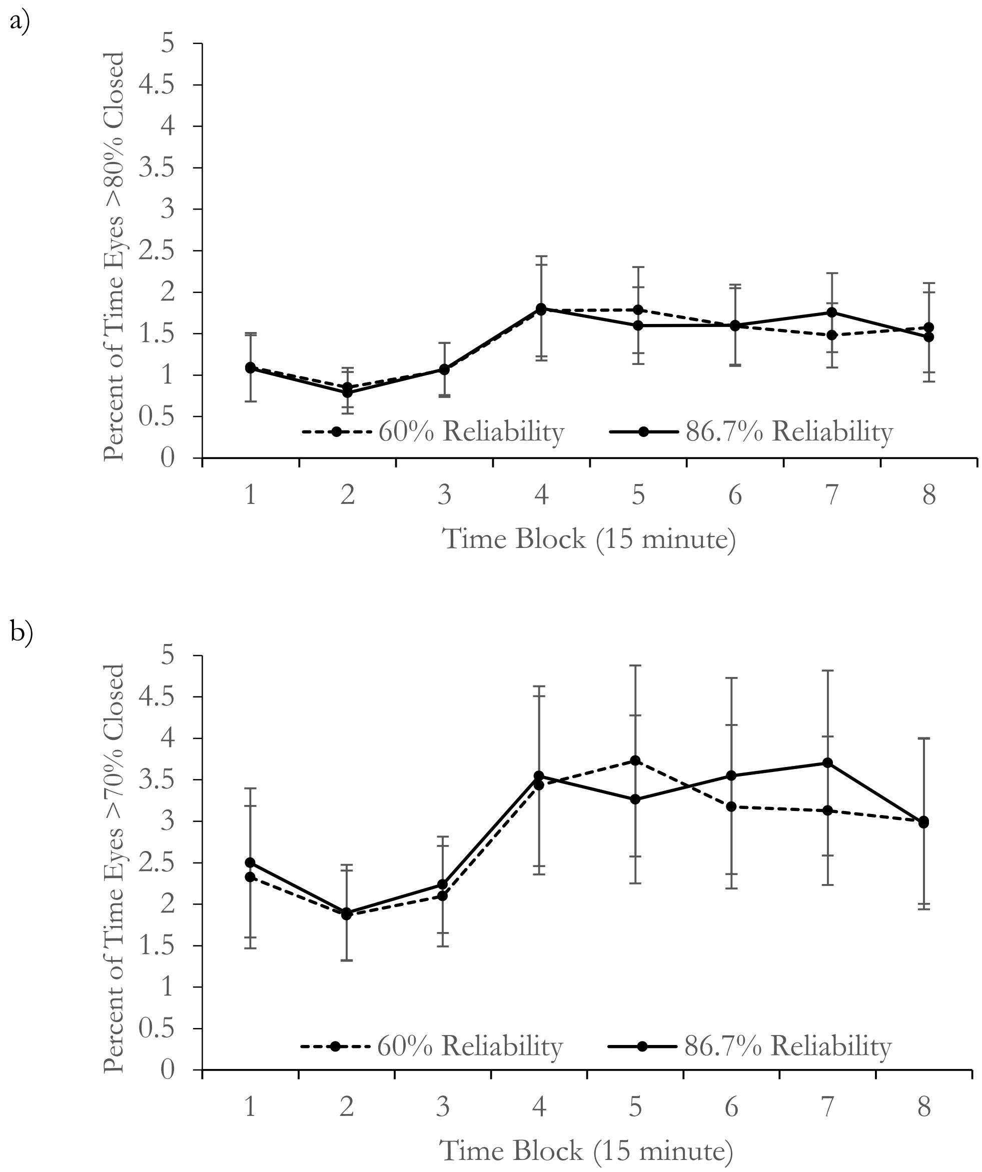


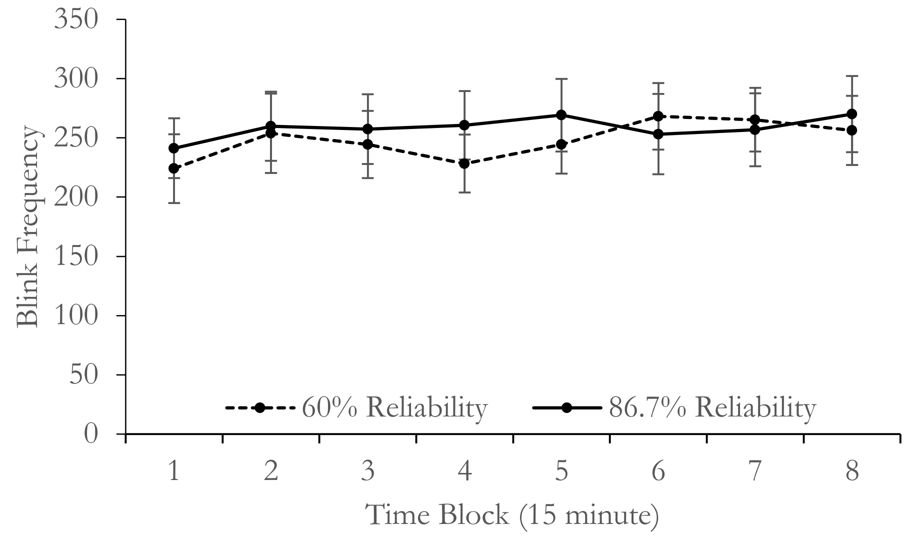
Figure 11. Percent of time eyes were more than 80% (a) and 70% (b) closed. Bars represent standard error.

Blinks

Frequency. Mauchly's test indicated a violation of sphericity, $\chi^2(27) = 76.53.00$, p < .001. Degrees of freedom were corrected using Box's estimates of sphericity (ε = .56). Results revealed no time or reliability effect on blink frequency (See Figure 12).

Duration. Blink duration data did not violate the assumption of sphericity, $\chi^2(27) = 38.80$, p = .068. No effects were significant at the $\alpha = 0.05$ level. However, there was a marginal finding for time on task, $F(7, 259) = 1.98$, $p = .058$, $\eta^2_p = .051$, which planned trend analysis indicated was quadratic in nature, $F(1, 37) = 5.57$, $p = .024$, $\eta^2_p = .131$. Blink duration increased at the beginning of the mission, leveled off, and then began to decline toward the end (Figure 12).

a)



b)

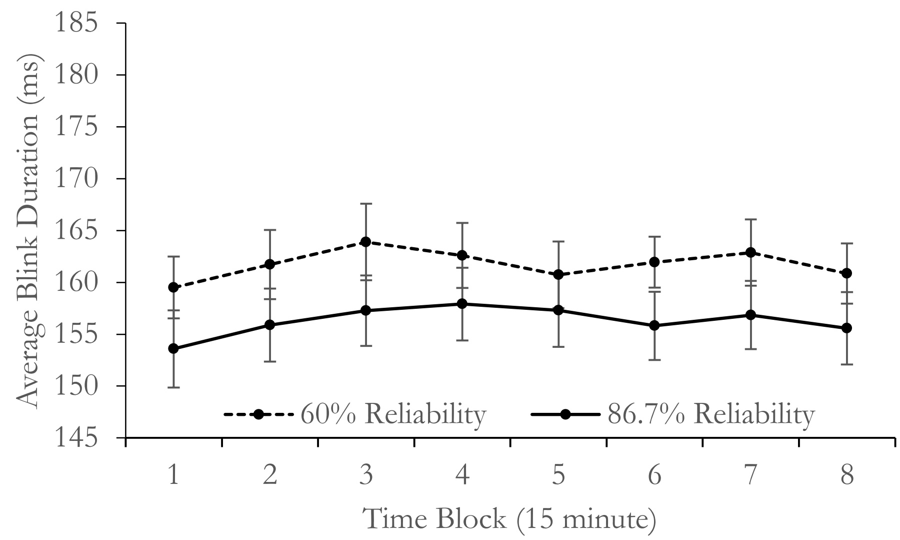


Figure 12. Blink frequency (a) and average duration (b). Bars represent standard error.

Fixations

Mean Fixation Duration. Mauchly's test indicated a violation of sphericity, $\chi^2(27) = 65.24$, p < .001. Degrees of freedom were corrected using Box's estimates of sphericity ($\varepsilon = .64$). There was a main effect for time $F(4.51, 153.28) = 2.54$, $p = .036$, $\eta^2_p = .070$, which planned analysis indicated was driven by a quadratic trend, $F(1, 34) = 7.39$, $p = .010$, $\eta^2_p = .178$. Fixation duration generally

declined until period three or four, and then slowly increased with time on task (Figure 13). A marginal, but not significant, $F(1, 34) = 3.24$, $p = .081$, $\eta^2_p = .087$, effect for reliability suggested that fixations were longer on average in the low reliability condition.

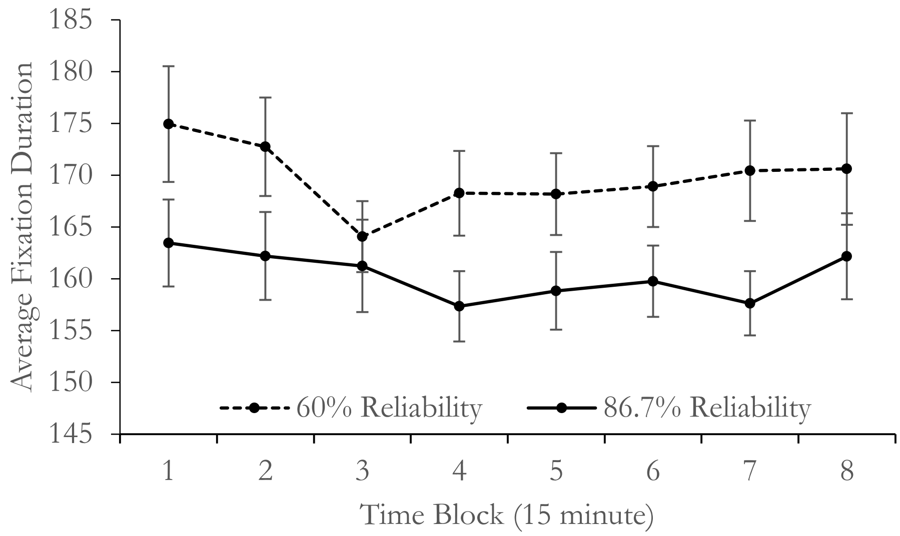


Figure 13. Average fixation duration. Bars represent standard error.

Express Fixation (<150 ms) Frequency. Mauchly's test indicated a violation of sphericity, $\chi^2(27) = 107.86$, p < .001. Degrees of freedom were corrected using Box's estimates of sphericity ($\varepsilon = .45$). There was a strong main effect for time, $F(3.18, 108.06) = 24.55$, $p < .001$, $\eta^2_p = .419$. Planned trend analysis indicated that this effect was driven by both linear $F(1, 34) = 38.90$, $p < .001$, $\eta^2_p = .534$ and quadratic trends $F(1, 34) = 13.13$, $p = .001$, $\eta^2_p = .279$. Generally, express fixation frequency decreased with time on task. However, there was a slight increase at the beginning of the mission and a sharp decline at the end of the mission (see Figure 14). There was also a marginal interaction effect, $F(3.18, 108.06) = 2.19$, $p = .090$, $\eta^2_p = .060$. A visual inspection suggests that the differences between reliability condition in the beginning of the mission attenuated with time.

Cognitive Fixation (150 ms – 900 ms) Frequency. Mauchly's test indicated a violation of sphericity, $\chi^2(27) = 90.60$, p < .001. Degrees of freedom were corrected using Box's estimates of sphericity (ε = .53). Results showed a main effect for time, $F(3.70, 126.04) = 15.88$, $p < .001$, $\eta^2_p = .318$, with planned trend analysis indicating a linear trend, $F(1, 34) = 32.10$, $p < .001$, $\eta^2_p = .486$. Average cognitive fixation frequency decreased with time on task (Figure 14*Figure 14*. Frequency of express (<150 ms; a), cognitive (150-900 ms; b), and overlong (>900 ms; c) fixations. Bars represent standard error.).

Overlong Fixation (>900 ms) Frequency. Mauchly's test indicated a violation of sphericity, $\chi^2(27) = 84.22$, p < .001. Degrees of freedom were corrected using Box's estimates of sphericity (ε = .55). There were no significant findings for overlong fixations at the α = 0.05 level. However, there was a marginal main effect for time on task, $F(3.82, 130.03) = 1.98$, $p = .058$, $\eta^2_p = .055$, which planned analysis suggested was driven by a linear trend, $F(1, 34) = 3.43$, $p = .073$, $\eta^2_p = .092$. Overlong fixations decreased with time on task (Figure 14).

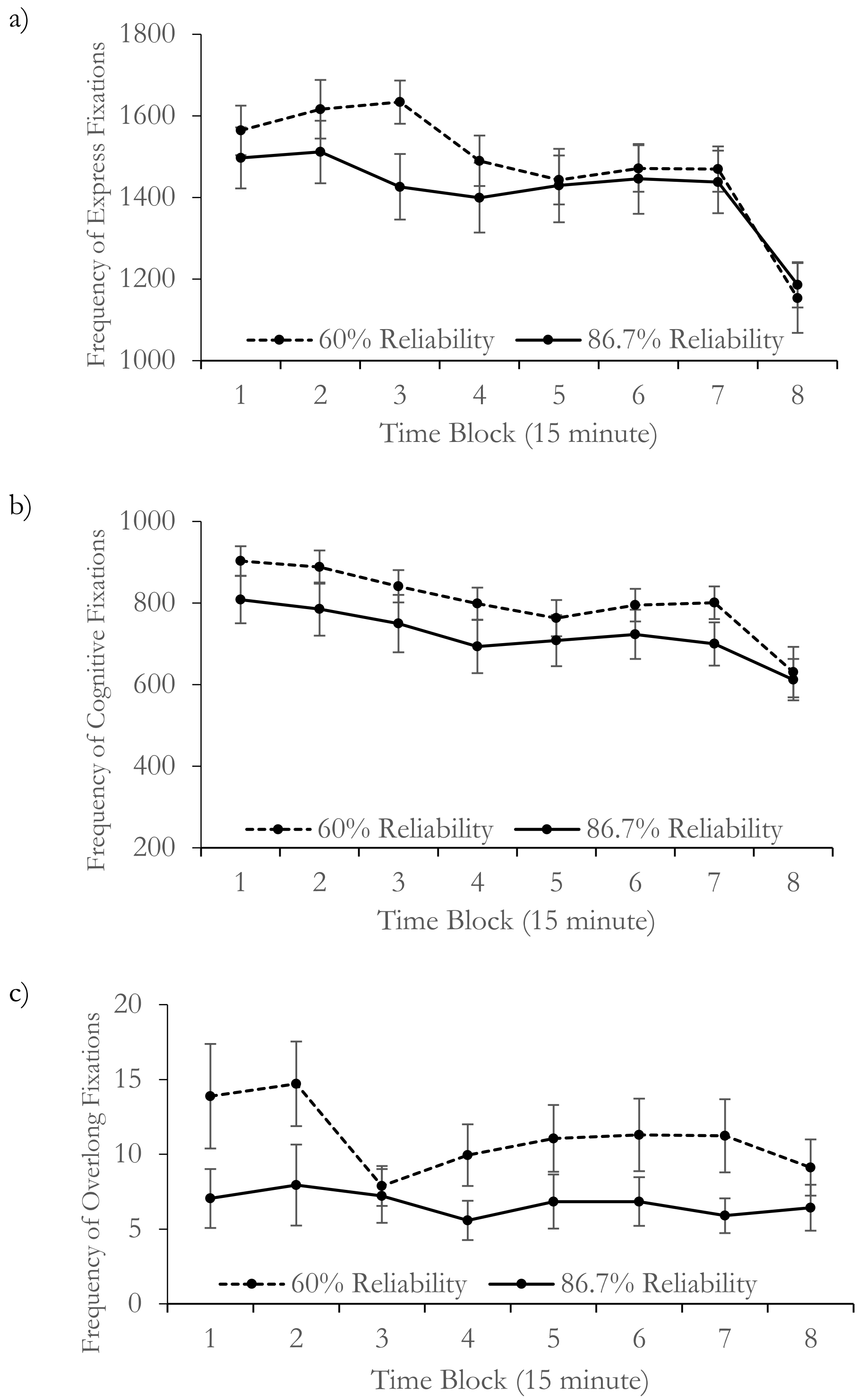


Figure 14. Frequency of express (<150 ms; a), cognitive (150-900 ms; b), and overlong (>900 ms; c) fixations. Bars represent standard error.

Percent Cognitive Fixations. Among the fixation bins, only percentage of cognitive fixations were chosen for analysis because they were the focus of previous theoretically driven work (Schleicher et al., 2008) and as well as for practical reasons. Overlong fixations were negligible and therefore express fixations are nearly directly dependent on their proportion to cognitive fixations, making their analysis unnecessary (see relative frequency of each bin in Figure 14).

Mauchly's test indicated a violation of sphericity, $\chi^2(27) = 59.33$, $p < .001$. Degrees of freedom were corrected using Box's estimates of sphericity ($\varepsilon = .62$). There was a main effect of time on task for percent of fixations which were in the cognitive range, $F(4.31, 146.52) = 3.62$, $p = .006$, $\eta^2_p = .096$. Planned follow-up analyses indicated a quadratic trend, $F(1, 34) = 11.40$, $p = .002$, $\eta^2_p = .251$, and a marginal linear trend, $F(1, 34) = 3.48$, $p = .071$, $\eta^2_p = .093$. Relative cognitive fixation occurrence decreased at the beginning of the mission but then began to increase steadily after the third or fourth period (Figure 15).

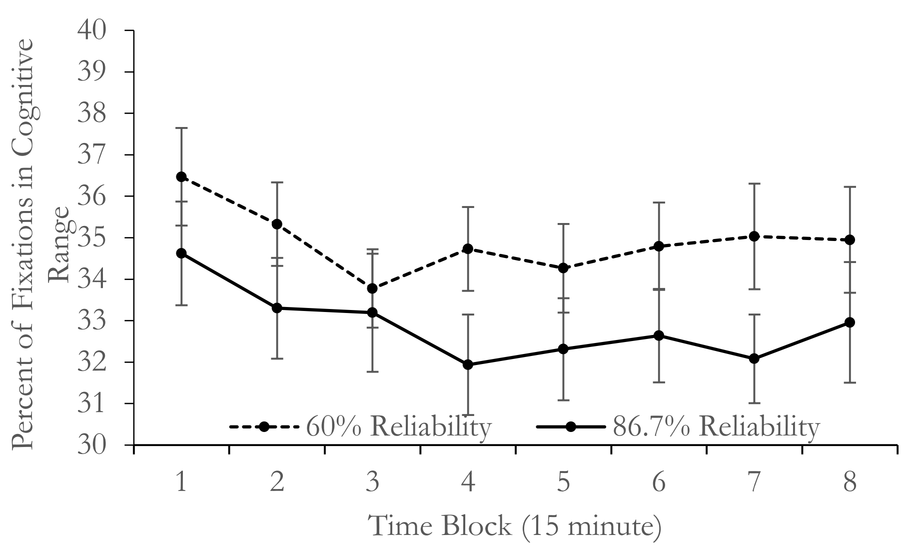


Figure 15. Percent of fixations that within the cognitive range (150-900 ms). Bars represent standard error.

Percent Determinism (RQA)

RQA analysis is sensitive to data quality issues and requires higher data quality than I was able to obtain in this effort. Therefore, there are no RQA outcomes to report.

Dwell Time Based Reliance

Total dwell time in different areas of interest (AOI), was used to infer reliance on the automated decision making aid. Specifically, I assumed that for the image analyses tasks, the less time spent looking at the answer portion of the task relative to the image, the more the participant relied on the automation. Therefor the calculation of this metric (Equation 1) is the amount of time spent gazing at the answer subsection of the image analysis window, area 1.2 (a), compared to total dwell time within the window – areas 1.2 (a) and 1.1 (b; see Figure 16).

$$Eye\ Tracking\ Reliance = a/(a + b) \tag{1}$$

Figure 16. Subsections of the image analysis window.

For this analysis, Mauchly's test indicated a violation of sphericity, $\chi^2(27) = 79.37$, p < .001. Degrees of freedom were corrected using Box's estimates of sphericity ($\varepsilon = .59$). Results indicated no effect for automation reliability, but did indicate variation over time (see Figure 17), $F(4.15, 153.44) = 5.77$, $p < .001$, $\eta^2_p = .135$. Planned trend analysis implicated a small quadratic trend, $F(1, 37) = 5.50$, $p = .025$, $\eta^2_p = .129$. There was no significant interaction.

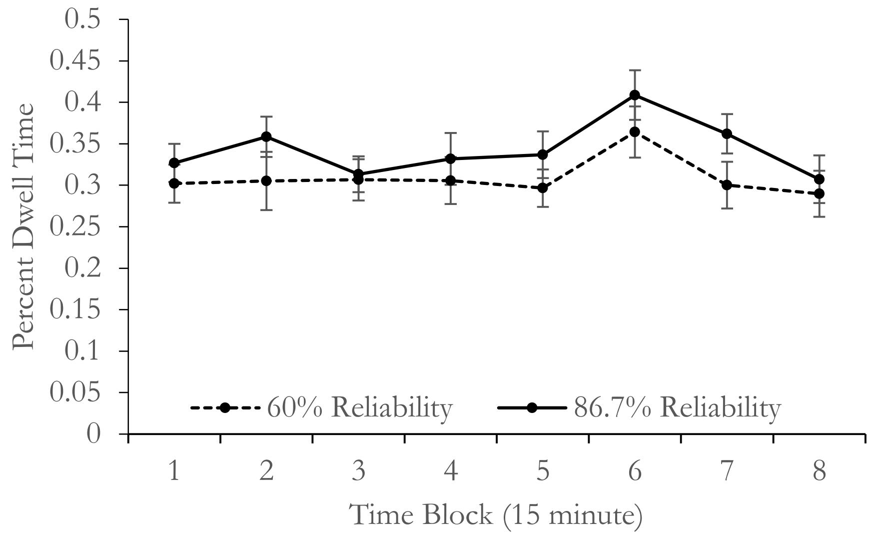


Figure 17. Percent of Image Task window dwell time spent within the answer section. Higher percent suggests greater reliance on automation. Bars represent standard error.

Correlations between Subjective Fatigue and Performance-Based Outcomes

Table 4 lists correlations between post-task stress state factors and performance outcomes. The eighth period was used as a correlate because the DSSQ asked participants to rate their states the very end of the task. Additional correlational analyses were conducted for the sixth and seventh periods because participants may have changed their performance regulation strategies toward the conclusion of the task, allocating increasingly more effort toward in anticipation of the end of the task, a phenomenon called the end spurt (Bergum & Lehr, 1963). This concern was supported in the current effort by quadratic trends in weapons release task performance and several eye tracking metrics. For this analysis, there was a significant positive relationship between engagement and accuracy for the Image Analysis task across reliability condition. Engagement also predicted performance in the sixth period of the Weapon Release Authorization task.

Table 4

Correlations between Engagement and Performance Metrics for Period 6, 7, and 8

	6th Period		7th Period		8th Period	
Variable	*r*	*p*	*r*	*p*	*r*	*p*
Accuracy						
Image Analysis	.23	.008	.23	.007	.10	.257
Weapon Release Authorization	.17	.048	.07	.454	.12	.176
Reliance						
Image Analysis	.00	.970	.05	.573	.10	.257
Weapon Release Authorization	.15	.084	.04	.665	.12	.176

Because of the possibility that engagement plays a different role depending on the reliability of the automation, a series of correlations was run for each reliability condition individually (Table 5). These correlations indicated that engaged participants generally performed better in both tasks. Notably, engaged participants showed a nonsignificant trend towards lower reliance on the low reliability automation for the Image Analysis task but more reliant on the low reliability automation for the Weapon Release Authorization task.

Table 5

Correlations (r) between Engagement and Performance Metrics for Period 6, 7, and 8 by reliability

	60% Reliability			86.7% Reliability		
Variable	6	7	8	6	7	8
Accuracy						
Image Analysis	.27*	.27*	.07	.17	.17	.13
Weapon Release Authorization	.21	-.01	.12	.12	.09	.12
Reliance						
Image Analysis	-.24	.01	-.09	.19	.02	.10
Weapon Release Authorization	.25*	.00	-.02	.03	.01	-.01

* $p < .05$, ** $p < .01$

Correlations between Subjective Fatigue and Eye Tracking Metrics

Correlations between post-task task engagement and each of the last three periods for the candidate eye tracking metrics were computed to determine whether objective metrics converged with subjective fatigue (Table 6). Blink duration, mean fixation duration, frequency of express and overlong fixations, and percent of fixations occurring within the cognitive time range all predicted task engagement in one or more of the final periods. Blink frequency also showed a consistent but weaker relationship with task engagement.

Table 6

Correlations between Engagement and Eye Tracking Metrics

	6th Period		7th Period		8th Period	
	r	*p*	*r*	*P*	*r*	*p*
PERCLOS 80%	-.14	.391	-.08	.644	-.18	.261
PERCLOS 70%	-.06	.741	-.01	.966	-.15	.366
Blink Frequency	-.25	.127	-.25	.124	-.22	.177
Blink Duration	-.32	.049	-.28	.090	-.23	.158
Mean Fixation Duration	-.43	.009	-.43	.010	-.26	.118
Express Fixation Frequency	.30	.078	.22	.191	.44	.007
Cognitive Fixation Frequency	-.04	.811	-.13	.444	.20	.242
Overlong Fixation Frequency	-.39	.018	-.37	.026	-.17	.312
Percent Cognitive Fixations	-.36	.032	-.36	.032	-.19	.274
Eye Tracking Reliance	.03	.872	.21	.198	.09	.592

It was not reasonable to split the data by reliability condition for separate analysis as was done for analysis for engagement and performance outcomes due to the low eye tracking sample size. Further, it is important to note that because of this small sample size for these eye tracking-based correlations and those that follow, power is limited, and small to medium effect patterns may be difficult to uncover.

Correlations between Performance-Based Outcomes and Eye Tracking Metrics

To understand the relationship between accuracy, reliance, and the candidate eye tracking metrics, a series of bivariate correlations was run. Correlations were not run separately by automation reliability because it would be impossible to gauge these difference conditions in practice.

Accuracy

Image Analysis. Image Analysis accuracy correlated with several eye tracking metrics in the second half of the mission, particularly in the sixth and seventh periods (Table 7). Blink duration, fixation duration, and percent of fixations occurring within the cognitive time range, all had multiple periods predict Image Analysis accuracy.

Table 7

Correlations (r) between Image Analysis Task Accuracy and Eye Tracking Metrics

	Period							
	1	2	3	4	5	6	7	8
PERCLOS 80%	-.17	-.01	-.09	-.26	-.05	-.08	-.32*	-.23
PERCLOS 70%	-.19	.00	-.02	-.23	-.01	-.08	-.35*	-.27
Blink Frequency	-.10	-.26	.11	-.01	-.14	-.02	-.29	-.08
Blink Duration	-.28	-.09	-.04	-.20	-.33*	-.20	-.53**	-.31
Mean Fixation Duration	-.22	.00	-.30	-.19	-.16	-.44**	-.50**	-.08
Express Fixation Frequency	.07	-.23	-.17	-.16	.21	.01	.16	.34*
Cognitive Fixation Frequency	-.17	-.22	-.26	-.28	.05	-.30	-.18	.16
Overlong Fixation Frequency	-.20	.07	-.21	-.06	-.14	-.23	-.44**	-.08
Percent Cognitive Fixations	-.25	-.08	-.26	-.29	-.10	-.45**	-.40*	-.07
Eye Tracking Reliance	-.06	.02	-.11	-.26	.16	-.32*	-.07	-.01

* $p < .05$, ** $p < .01$

Weapon Release Authorization. Eye tracking metrics did not seem to predict weapon release accuracy well. Those correlations that did occur may be spurious or at least unhelpful (Table 8). Express and cognitive fixation frequency were each positively correlated with accuracy in the second

period but then the relationship fluctuated between near zero correlations and negative correlations. With the exception of the second period, overlong fixation frequency had consistent negative relationships with accuracy, including a significant negative correlation in the fifth period.

Table 8

Correlations (r) between Weapon Release Authorization Task Accuracy and Eye Tracking Metrics

	Period							
	1	2	3	4	5	6	7	8
PERCLOS 80%	-.10	-.11	-.05	.07	.25	.04	.02	-.11
PERCLOS 70%	-.12	-.14	-.02	.08	.26	.08	.06	-.08
Blink Frequency	-.18	-.10	-.09	-.19	.02	-.08	.04	-.24
Blink Duration	.01	.02	-.12	-.04	.21	.13	-.01	.07
Mean Fixation Duration	-.18	.07	.16	-.06	-.13	-.26	-.16	-.10
Express Fixation Frequency	.15	.42*	.30	-.10	-.03	-.20	-.18	.03
Cognitive Fixation Frequency	-.06	.34*	.21	-.10	-.01	-.27	-.15	.03
Overlong Fixation Frequency	-.15	.08	.15	-.03	-.34*	-.17	-.31	-.26
Percent Cognitive Fixations	-.17	.16	.11	-.05	.02	-.22	-.01	.01
Eye Tracking Reliance	.01	-.05	.05	.28	.07	.01	.09	-.01

* $p < .05$, ** $p < .01$

Reliance

Image Analysis. There were some correlations between eye tracking metrics and reliance on automation for the Image Analysis task. Although there was a potentially spurious significant negative correlation with express fixation frequency, significant correlations for blink duration, fixation duration, overlong fixations, and percent cognitive fixations, are more credible, as each of these metrics showed a consistent negative relationship with reliance through all periods (Table 9). However, the eye tracking reliance index was unrelated to the fatigue metrics.

Table 9
Correlations (r) between Image Analysis Task Reliance and Eye Tracking Metrics

	Period							
	1	2	3	4	5	6	7	8
PERCLOS 80%	.02	-.06	-.16	-.01	-.09	-.11	-.02	-.20
PERCLOS 70%	.04	.04	-.08	.00	-.11	-.09	-.12	-.22
Blink Frequency	.09	.19	.23	.11	-.01	-.04	-.27	-.10
Blink Duration	-.26	-.16	-.01	-.21	-.12	-.29	-.13	-.32*
Mean Fixation Duration	-.15	-.15	-.01	-.07	-.42*	-.31	-.24	-.20
Express Fixation Frequency	-.12	.05	-.39*	-.06	.03	.01	-.09	.05
Cognitive Fixation Frequency	-.19	.02	-.19	-.05	-.17	-.15	-.22	-.06
Overlong Fixation Frequency	-.17	-.22	.02	-.04	-.36*	-.30	-.31	-.26
Percent Cognitive Fixations	-.13	-.06	.00	-.07	-.34*	-.24	-.18	-.14
Eye Tracking Reliance	-.12	.14	.03	-.07	.10	.13	.08	-.08

$* p < .05$, $** p < .01$

Weapon Release Authorization. Correlations with Weapon Release Authorization reliance on automation was somewhat consistent with that for the Image Analysis task. Again blink duration, fixation duration, overlong fixations, and percent cognitive fixations showed consistent negative relationships. This was especially true for fixation duration and percentage of cognitive fixations, which each had two significant correlations and consistent negative relationships with reliance in general for the second half of the mission (Table 10).

Table 10

Correlations (r) between Weapon Release Authorization Task Reliance and Eye Tracking Metrics

	Period							
	1	2	3	4	5	6	7	8
PERCLOS 80%	.03	.10	.03	.19	.05	-.15	.04	-.15
PERCLOS 70%	.06	.16	.03	.20	.07	-.05	.09	-.13
Blink Frequency	-.03	.07	-.09	.01	-.04	.00	-.12	.15
Blink Duration	-.31	.02	-.22	-.12	-.17	-.21	-.17	-.07
Mean Fixation Duration	.21	-.20	.12	-.14	-.38*	-.39*	-.30	-.23
Express Fixation Frequency	-.22	.03	.08	-.15	-.07	.09	.04	-.07
Cognitive Fixation Frequency	.03	-.01	.14	-.14	-.25	-.17	-.24	-.22
Overlong Fixation Frequency	.12	-.24	.06	-.17	-.36*	-.33	-.10	-.14
Percent Cognitive Fixations	.22	-.08	.13	-.12	-.30	-.34*	-.37*	-.30
Eye Tracking Reliance	.09	.09	.08	.10	.14	.01	.30	-.07

* $p < .05$, ** $p < .01$

Prediction of Fatigue

Three hierarchical regressions were conducted to understand the importance of online performance- and eye tracking-based metrics for predicting fatigue online. Task engagement as measured by the DSSQ, the hypothesized primary indicator of passive fatigue, was the outcome variable for all three regressions. Depending on regression, the value for the sixth, seventh, or eighth period was used for each predictor variable. The eighth period was used because instructions for the DSSQ asked participants to report their fatigue toward the end of the task. As with the correlational analysis, assessment of the sixth and seventh period were conducted due to concerns of the impact of the end spurt phenomenon (Bergum & Lehr, 1963) on data recorded toward the end of the task. This final spurt may confound efforts to gauge accurately the level of fatigue induced by task factors.

For the first, simple model, performance-based reliance and PERCLOS using 80% eye closure were entered. For the second, complex model, percent fixations within the cognitive range was added to the model. The original intent was to include a determinism metric base on RQA also, but data quality issues prevented analysis using this metric. In its place, frequency of express

fixations was added to the second model, because it had the strongest relationship with time on task of the eye tracking metrics and because it fit with the case made for variables to be added as part of the complex model.

The simple first model was not significantly different from zero for periods six, $F(2, 35) = 0.41$, $p = .669$, seven, $F(2, 35) = 1.08$, $p = .353$, or eight, $F(2, 35) = 0.82$, $p = .449$. The complex model for the sixth (Table 11), $F(4, 35) = 2.64$, $p = .052$, and seventh (Table 12), $F(4, 35) = 2.10$, $p = .105$, periods were not significant at the $\alpha = 0.05$ level, but the marginal *p* values suggest that they may have been with more power. Percentage of cognitive fixations drove both models ($p = .015$ and $p = .037$, respectively) as did, to a lesser extent, the frequency of express fixations in the sixth period ($p = .075$).

For the eighth and final period, results revealed that *R* for the complex model was significantly different from zero, $F(4, 35) = 3.01$, $p = 033$. The model was driven by the frequency of express fixations metric (see Table 13). The percent of fixations within the cognitive range also may have contributed to the model ($p = .129$) had there been more statistical power.

Table 11

Hierarchical Regression of Eye Tracking Metrics on Passive Fatigue as Indicated by Subjective Task Engagement for Period 6

	Model 1			Model 2		
Variable	*B*	*SE B*	β	*B*	*SE B*	β
Dwell Time Based Reliance	2.88	7.76	.07	8.21	7.44	.19
PERCLOS 80% Closure	-0.45	0.51	-.16	-0.27	0.47	-.10
Percent Cognitive Fixations				-53.07	20.59	-.43*
Frequency of Express Fixations				0.01	0.00	.30
R^2		.02			.25	
ΔR^2		.02			.23	

* $p < .05$, ** $p < .01$

Table 12

Hierarchical Regression of Eye Tracking Metrics on Passive Fatigue as Indicated by Subjective Task Engagement for Period 7

	Model 1			Model 2		
Variable	*B*	*SE B*	β	*B*	*SE B*	β
Dwell Time Based Reliance	12.05	8.64	.24	8.77	8.31	.18
PERCLOS 80% Closure	-0.42	0.54	-.13	-0.32	0.52	-.10
Percent Cognitive Fixations				-40.17	18.48	-.35*
Frequency of Express Fixations				0.00	0.00	.19
R^2		.06			.21	
ΔR^2		.06			.15	

* $p < .05$, ** $p < .01$

Table 13

Hierarchical Regression of Eye Tracking Metrics on Passive Fatigue as Indicated by Subjective Task Engagement for Period 8

	Model 1			Model 2		
Variable	*B*	*SE B*	β	*B*	*SE B*	β
Performance Based Reliance	5.56	8.17	.12	4.41	7.72	-.09
PERCLOS 80% Closure	-0.50	0.43	-.20	-0.41	0.39	-.16
Percent Cognitive Fixations				-26.07	15.83	-.26
Frequency of Express Fixations				0.01	0.00	.44**
R^2		.05			.28*	
ΔR^2		.05			.23*	

* $p < .05$, ** $p < .01$

For regressions predicting task performance there were no significant findings for Weapon Release Authorization task accuracy. There were the following findings for Image Analysis task accuracy. The simple first models were not significantly different from zero for periods six, $F(2, 35) = 1.88$, $p = .168$, seven, $F(2, 35) = 1.85$, $p = .172$, or eight, $F(2, 35) = 0.94$, $p = .401$. The complex model for the sixth (Table 14), $F(4, 35) = 2.42$, $p = .069$, and eighth (Table 16), $F(4, 35) = 1.54$, $p = .215$, periods were not significant at the $\alpha = 0.05$ level, but the marginal p values suggest that the

model for period six may have been significant with more power. Percent of fixations within the cognitive range alone powered the sixth model (p = .025).

For the seventh period, R for the complex model was significantly different from zero, F(4, 35) = 3.14, p = 028. The model was driven mostly by the percent cognitive fixations metric (see Table 15). PERCLOS may have contributed to the model (p = .067) had there been more statistical power.

Table 14

Hierarchical Regression of Eye Tracking Metrics on Image Analysis Accuracy for Period 6

	Model 1			Model 2		
Variable	*B*	*SE B*	β	*B*	*SE B*	β
Dwell Time Based Reliance	-44.19	23.50	-.32	-25.86	23.76	-.19
PERCLOS 80% Closure	0.01	1.53	-.00	-0.10	1.51	-.01
Percent Cognitive Fixations				-154.37	65.73	-.39*
Frequency of Express Fixations				0.00	0.01	.03
R^2		.10			.24	
ΔR^2		.10			.14	

* $p < .05$, ** $p < .01$

Table 15

Hierarchical Regression of Eye Tracking Metrics on Image Analysis Accuracy for Period 7

	Model 1			Model 2		
Variable	*B*	*SE B*	β	*B*	*SE B*	β
Dwell Time Based Reliance	1.23	19.88	.01	-6.09	18.58	-.05
PERCLOS 80% Closure	-2.34	1.24	-.32	-2.20	1.16	-.30
Percent Cognitive Fixations				-113.93	41.33	-.42*
Frequency of Express Fixations				0.01	0.01	.13
R^2		.10			.29*	
ΔR^2		.10			.19*	

* $p < .05$, ** $p < .01$

Table 16

Hierarchical Regression of Eye Tracking Metrics on Image Analysis Accuracy for Period 8

	Model 1			Model 2		
Variable	*B*	*SE B*	β	*B*	*SE B*	β
Performance Based Reliance	1.89	15.25	.02	-1.97	15.56	-.02
PERCLOS 80% Closure	-1.10	0.80	-.23	-0.94	0.78	-.20
Percent Cognitive Fixations				-19.16	31.90	-.10
Frequency of Express Fixations				0.01	0.01	.33
R^2		.05			.17	
ΔR^2		.05			.11	

* $p < .05$, ** $p < .01$

It should be noted that the sample size ($N = 39$) for these eye tracking-based multiple regressions is too small, and these results should be interpreted with caution.

DISCUSSION

Generally, the findings indicated that the mission elicited passive fatigue with time on task based on marked decreases in subjective task engagement. In addition, the low reliability automation condition elicited more subjective workload than the high reliability automation condition. Participants trusted the low reliability automated aid less on average. Performance outcomes indicated that accuracy and reliance were higher in the high reliability condition. Accuracy increased slightly for the Image Analysis task and decreased slightly for the Weapon Release Authorization task with time. A stronger quadratic pattern suggested that participants anticipated the end of the mission. Reliance on automation did not increase or decrease for the Image Analysis task, but decreased with time in both reliability conditions for the Weapon Release task. This decrease with time was more rapid for the low reliability condition.

All eye tracking metrics registered changes with time on task with the exception of blink frequency and the eye tracking measure of reliance. Additionally, the mean fixation duration registered differences by reliability condition. Only express fixation frequency showed any sign of an interaction between reliability and time on task but it did not support the hypothesized increase in fatigue discrepancy between conditions with time on task. Regressions were conducted to predict subjective task engagement and performance as a test of the goal to create fatigue detection algorithms. The complex algorithms were successful at predicting both task engagement and accuracy, primarily on the strength of the metric for percentage of fixations within the cognitive time range. Standardized Beta values suggested that dwell time based reliance, 80% PERCLOS, and express fixation frequency may also contribute consistently to detection were there more power to verify their effect. However, regression findings are tentative because of the limited sample size for analysis for eyetracking data. All of the above findings are elaborated below in the order of the specific aims.

Evidence for Fatigue

Before addressing the aims, it is important to determine whether fatigue was induced by the mission and if so, what kind of fatigue it is likely to be. The first step in this determination is to examine the task characteristics. The NASA-TLX indicated that across both tasks, temporal demand, effort, and frustration were below the midpoint workload level of 50. Mental workload was just above the middle value in the low reliability condition, but below the medium value with the reliable automated aid. Overall, the total workload was M = 39.65 in the low reliability condition and M = 32.26 in the high reliability condition. For comparison, a very similar experiment which used the same simulator and tasking for a one hour mission (Lin et al., 2015) had total workload levels of M = 57.0 for a high workload manipulation and M = 46.2 for a low workload condition. Thus the low workload conditions implicated in passive fatigue (Desmond & Hancock, 2001; Saxby et al., 2013) were present in the mission.

The most valid and discriminating gauge of fatigue in this effort was the DSSQ which served as a benchmark for evaluating other measures. Previous work (Saxby et al., 2007) has found that active fatigue is marked by an increase in distress and passive fatigue is associated with a decrease in task engagement. In the current effort, distress was unchanged and task engagement decreased considerably. On this evidence, it can be concluded that the mission elicited passive fatigue. However, there is more evidence to reinforce this conclusion. Passive fatigue has been associated with vigilance decrement in simulated driving tasks supported by various levels of automation (Körber, Cingel, Zimmermann, & Bengler, 2015; May & Baldwin, 2009; Schmidt et al., 2009; Thiffault & Bergeron, 2003). Therefore, another way to test for passive fatigue would be to look for a performance decline in mission tasking.

The loss of task engagement gauged by the DSSQ and the performance decrement recorded for the Weapon Release task supported the hypothesis that the simulated UAV mission would

induced passive fatigue. If indeed, passive fatigue induced an energy conservation approach (Sauer et al., 2003), the null finding for the Image Analysis task might be explained if participants found completing it to be less effortful than completing the Weapon Release Authorization task. Post study interviews suggest that tanks were quite difficult to identify.

Evaluation of Candidate Metrics for Gauging of Operator Fatigue

A primary goal of this effort was to identify measures that could be used to detect fatigue on a continuous basis during performance. Ideally, these measures would meet selection criteria for sensitivity, diagnosticity, intrusiveness, and robustness (Eggemeier, Wilson, Kramer, & Damos, 1991; Wohleber, Matthews, Funke, & Lin, 2016). Sensitivity can refer either to signal-to-noise ratio or the quickness with which a measure can detect state changes. Concerning the first, it is important to choose a metric that is sensitive within the range of expected signals, but not necessarily one that is sensitive over the entire range of possible signals. The present effort focused on variance in fatigue while operators were relatively wakeful so that severe fatigue could be prevented from occurring altogether. Therefore, by the sensitivity criteria for this effort, a measure that is very sensitive to variations at higher levels of fatigue (drowsiness), but relatively insensitive to variation at lower levels of fatigue, was not ideal. The present research design did not afford an analysis of signal latency; however, it is suspected that signal latency values would be much smaller than the time unit of analysis (15 minutes), and that time insensitivity of even several minutes would still prove useful and provide useful results in this effort.

Diagnosticity is the ability of a measure to distinguish between different aspects of a construct. It is especially important for a measure of fatigue, which is multifaceted and can elicit a range of behavioral outcomes different ways. One example outlined in the introduction is the different behaviors of drowsiness and DWA-like states, both symptoms of passive fatigue. Ideally, metrics for online detection of fatigue would account for both situations. In practice, optimal

diagnosticity of multiple outcome types may be possible only through joint monitoring using multiple metrics.

Intrusiveness concerns the level of disruption caused by the use of a measure. For example, despite the proven potential of EEG for online detection of fatigue (Siegmund et al., 1996), it is relatively impractical for use in day to day operations supported by closed-loop adaptive systems in UAV operation. For example, the setup of a capable EEG system would require an assistant to place electrodes, baseline calibration, and pre-task testing (e.g., impedance checking). During the shift, occasional calibration checks would have to be included to correct electrode drift and/or unsecured electrodes. The cords attached to the equipment, and the need not to disrupt the equipment setting might restrict operator motion. The headpiece participants might need to wear could become uncomfortable after long periods. Operators would also have to dismantle, clean, and store the equipment.

Robustness refers to the need for measures and metrics to be resilient to quality problems, such as a noisy operating environment, varying conditions, and equipment problems. There are three considerations. The first consideration is whether missing data for a metric tends to be at random or systematic. If data is missing at random, a measure might provide only 500 out of 1000 possible observations, but nonetheless provide a reliable assessment. However, if loss of data is related to particular events in the scenario, the assessment is much more problematic. The second consideration is how easily the metric loses its ability to detect a state with decreasing data quality. In the present study the questions is whether the eye tracking metric depends on quality levels that are unsustainable for outcomes from the metric to be meaningful. The third and the potentially most challenging situation is whether the outcome indicated by the metric changes qualitatively as a result of shifts in quality of the data.

Although a more complete effort would need to be undertaken to truly validate metrics for online fatigue detection, this effort took a first step toward evaluating candidates for such efforts. Sensitivity was determined by the degree to which measures registered changes with time on task, and differences between conditions of automation reliability. For diagnosticity, alternative forms of fatigue were not assessed for discriminant validity. However, we did evaluate whether or not each measure could detect the specific facet of interest, passive fatigue. Thus, diagnosticity was evaluated based on the correlation of measures with subjective task engagement. Intrusiveness was not evaluated in this study but instead was used as a selection criterion for measures that would be evaluated in this effort. Eye tracking, for reasons outlined in the introduction, is unintrusive relative to other physiological measurement techniques, such as EEG, which might be used to gauge fatigue online. Finally, for any measure to prove useful by the analyses in this study, it must be robust. This effort aimed to recreate some of the realism of a real multi-UAV mission. Measures that detect fatigue must be robust to the presence of several different tasks, changes in event rate, and changes in the reliability of the automation. Further, robustness is especially prominent in this effort given the eye tracking quality issues. Metrics which produce inconclusive results may simply lack resilience to eye tracking quality issues, which is instructive nonetheless.

PERCLOS

Work using PERCLOS in passive fatigue situations has found that increases in eye closure predict vigilance task performance (Abe et al., 2011; Dinges et al., 1998). PERCLOS is typically used to detect conditions of severe fatigue (drowsiness) such as when a person is sleep deprived (e.g. McKinley et al., 2011), and is apt to exhibit lapses in attention and microsleeps (Abe et al., 2011; Dinges et al., 1998; Schleicher et al., 2008). Although PERCLOS has a well-established relationship with fatigue, it was unclear whether it would be sensitive to the range of fatigue present in the mission used in the present study. PERCLOS is typically used to predict catastrophic failure in

attention whereas the goal of fatigue detection for UAV operation is to optimize reliance on automation. The mission in the present study lasted only two hours after no sleep deprivation. Early work that attempted to use PERCLOS to classify fatigue states found difficulty classifying intermediate levels of fatigue with PERCLOS (<50% accuracy; Knipling & Wierwille, 1994).

Despite reservations, the results of this study found that both 80% and 70% versions of PERCLOS increased with time on task, as early work would predict. This increase had a medium effect size (η^2_p = .117-.133). Despite significant linear and quadratic trends, the abrupt increase in PERCLOS between the second and third period in both PERCLOS metrics followed by a leveling off suggest that changes in PERCLOS were not gradual. Instead, participants seemed to shift into a drowsy state after a half hour or so into the task. To the author's knowledge there is no investigation of PERCLOS which takes the same continuous time-based approach to PERCLOS as the present study with the exception of Kozak et al. (2005), who took hourly averages for three hours for participants performing a driving simulation task, and Funke (2011) who measured change in 10-minute increments during a vigilance task. Kozak and colleagues found a consistent increase of roughly 5% with each successive hour in sleep deprived drivers. It is unclear whether these increases occurred gradually or abruptly. However, Funke found an abrupt increase in PERCLOS during a stimulus task, after 30 minutes.

Notably, both PERCLOS metrics registered changes with time despite values remaining well within the "awake" (<7.5%) range with 80% PERCLOS between .5% and 2.5% and the 70% PERCLOS between 1% and 5%. For comparison, Kozak and colleagues' (2005) 80% PERCLOS average for sleep deprived participants in their first hour of driving was 3.88%, was 8.64% for the second hour, and 12.96% for the third hour. Despite this sensitivity, there was no difference between reliability conditions. The relatively low levels of fatigue may partially explain the lack of sensitivity to reliability.

Finally, both 80% and 70% PERCLOS were significantly correlated with Image Analysis task accuracy in the seventh period, and nearly significantly in the fourth and eighth period. It is logical that PERCLOS should correlate most strongly when fatigue is likely to be the greatest as in the seventh and eighth period. What is curious is that the fourth period should also show a sizable correlation. However, the fourth period was the period of abrupt increase in levels of PERCLOS. Interestingly, PERCLOS was not correlated with subjective task engagement. Nonetheless, PERCLOS appears to be a promising gauge of fatigue capable of showing changes in fatigue with time even at relatively low levels of drowsiness and useful for identifying poor performance.

Blinks

Previous work indicates that blink frequency and duration should increase with time on fatiguing tasks (Körber et al., 2015; J. A. Stern, Boyer, & Schroeder, 1994). According to Schleicher and colleagues (2008), blink frequency is more sensitive to lighter levels of fatigue and blink duration more sensitive to high levels. However, Körber et al. found both to increase in a passive fatigue lasting only 45 minutes, with changes in blink duration occurring prior to changes in blink frequency. In the present study, there was no effect of time on task or reliability for blink frequency. For blink duration there was a non-significant but marginal effect of time on task for blink duration explained by a quadratic trend similar to that of Körber et al. (2015), though the effect size was much lower and not significant despite both using a faceLAB Eye Tracker. One possibility is that the task in the present study was less fatiguing than the monotonous simulated drive employed by Körber et al. This explanation is supported by the fact that blink frequency and duration averages were lower for the present study than for Körber et al. Although there was no effect of blink duration, all the means for the low reliability condition were greater than means for the high reliability condition. This suggests that analysis with a bigger sample size might reveal a small effect of reliability.

Finally, higher blink duration was associated with poorer accuracy on the Image Analysis task throughout the mission. Correlations were especially strong in the second half of the mission and much stronger in general than that of PERCLOS. These findings are in line with previous work which showed that blink duration was negatively correlated with performance on sustained attention tasks (Morris & Miller, 1996; Verwey & Zaidel, 2000). Blink duration was also correlated with subjective task engagement, which suggests it is a valid gauge of passive fatigue. It is encouraging that blink duration is related so strongly to accuracy on the Image Analysis task and task engagement, but for an online metric to be effective it must be able to detect changes in fatigue state, which neither metric was able to do effectively.

Fixations

Mean Fixation Duration. Fixation duration can be impacted by many things including the prominence of daydreaming in vigilance tasking where participant become increasingly disengaged (Holmqvist et al., 2015). Work regarding fatigue and mean fixation duration has produced mixed outcomes. Some have found that mean or median fixation duration increases with time on passive fatigue inducing tasking (Funke, 2011; J. A. Stern, Boyer, Schroeder, Touchstone, & Stoliarov, 1994). Funke found that the positive association between mean fixation duration length and vigilance decrement occurred in a condition where there was certainty as to where the next signal would occur but not when there was spatial uncertainty. Other efforts have found no relationship between time on task and mean fixation durations (Lavine, Sibert, Gokturk, & Dickens, 2002; Schleicher et al., 2008).

Based on previous evidence, it was anticipated that fixation duration would increase with time on task. Instead, the findings reveal a significant change with time that took a quadratic pattern, dipping at the beginning, leveling out after the third or fourth period, then increasing steadily until the end of the task. Revisiting previous findings, one possible explanation could be the uncertainty

of the location of each new task within the present mission which may have prevented an increase in fixation duration as it did in Funke's effort (2011). Participants were tasked with monitoring the interface for seven different event types, occurring across several different windows. Additionally, the red plane task required participants to monitor the interface for a signal that could appear anywhere on a map which took up most of the right screen.

Task factor may partially explain why there the current findings for mean fixation duration and overlong fixation frequency are generally opposite of expectation. Stern et al. (1994) noted that some tasking may require longer fixations durations for successful task performance. In the present study, the mission required fixation on small points on several tasks. Participants had to identify highly degraded tanks images, and needed to read values from small displays to answer questions from a mission commander (e.g., report the heading of a vehicle). Participants were instructed not to move closer to the screen to perform these task, as this would move them outside of the view of the eye tracker. Fixations duration may be indicative of both increasing efficiency in identifying task elements, or of decreased effort in properly identifying elements with the onset of fatigue. Another explanation for the shortening of fixations might be that participant increasingly looked at locations where eye tracking was poor with increasing time on task. As participants got bored it is possible that they daydreamed more (Holmqvist et al., 2015) and during this daydreaming may have looked down at the keyboard or in other locations where poor tracking might have truncated fixation duration values. Nonetheless, correlational outcomes support the idea that longer fixations are to be expected with fatigue. In the mission, fixation duration was associated with lower task engagement and poorer accuracy for both the Image Analysis and Weapon Release Authorization tasks. Higher fixation duration was also related to lower reliance on automation in both tasks.

Binned Fixation Outcomes. The use of methods using comparison of binned fixation duration are new and apparently untested outside of the Schleicher et al. (2008) study on which my method is

based. Schleicher et al. found that the percentage of cognitive fixations (150 ms to 900 ms) relative to express (<150 ms) and overlong (>900 ms) fixations was positively correlated with self-rated alertness. They also found that expressed and overlong fixations greater for those who indicated low alertness. Other work has also chosen to assess the frequency of different ranges of fixation. Stern et al. (1994), for example, found that fixations with a duration longer than two seconds increased with time on a vigilance task.

In line with Schleicher et al.'s (2008) findings, there was a significant effect for percentage of fixations falling within the cognitive range with time on task. This was not a strong effect and was driven by a quadratic trend which may suggest that it was impacted by an end spurt. It is notable that this effect was present despite findings counter to expectations outlined below, which suggests that at a group level, percent cognitive fixations may be a robust metric. In this effort there was a main effect for express fixations and cognitive fixation frequency, but no effect for overlong fixations (which had a marginal effect indicating a slight decrease with time on task). The decrease in express fixations was unexpected but might be explained partially by difficulties with eye tracking quality, which is discussed in the next section. However, taking the decrease in express fixations at face value, there are other theoretically driven explanations.

Velickkovsky and colleagues have conducted several studies (e.g., Velichkovsky, Rothert, Kopf, Dornhöfer, & Joos, 2002) outlined in Velichkovsky, Joos, Helmert, and Pannasch (2005) which support a model of the visual system driven by two systems which impact eye movements. Specifically, there are two segments of fixation duration: one between 90 ms and 150 ms, which is thought to indicate an "ambient" mode of processing that involves orienting, and another of longer fixations, which are thought to relate to focal processing. Ambient mode is thought to relate roughly to nonconscious behavioral control of visuomotor behavior. A decrease in express fixations may indicate a growing familiarity with the visual scene. Further, this process of familiarity may be more

gradual in processing of dynamic visual information as new information is constantly being introduced. The interface used in the present mission is much more complex and unfamiliar than many used in eye tracking tasks investigating vigilance. It is possible that the rather abrupt decrease in period three and four reflect the point at which understanding of the visual properties of the interface is relatively complete at the unconscious level.

Correlational outcomes further complicate interpretation of outcomes. Overlong fixations were negatively correlated with task engagement and accuracy as was expected. Cognitive fixation frequency was not correlated with task engagement or significantly with accuracy. Express fixation frequency showed a pattern of correlation opposite to what was expected. Based on Schleicher et al.'s finding, express fixations should be indicative of decreased alertness. In the current study, express fixation frequency was positively correlated with task engagement and accuracy. In addition to the explanations above, this finding may be explained by improper binning caused by poor data quality, which is explained in more detail below.

Percent of Cognitive Fixations Paradox. One interesting problem that came up concerned the robustness issue of whether or not the outcome indicated by the metric changed qualitatively as a result of shifts in quality of the data. Specifically, an came up with the assessment of an eye tracking technique based on the binning of fixations by duration with guidance from a previous effort, which looked at alertness using EOG (Schleicher et al., 2008).

Our expectation was that with fatigue, participants would have a decreasing proportion of fixations between 150ms and 900ms, the purported cognitive fixation range, to total fixations. Although cognitive fixations declined as expected overall, we found that, at the individual level, our performance and subjective data suggested that those with the lowest proportion of cognitive fixations were the best performers and were the least fatigued. However, the method for identifying fixations we utilized was potentially sensitive to missing data, leading to an underestimation of

fixation duration. This meant that the quickest cognitive fixations may have been categorized as express fixations (<150 ms).

Cognitive fixations may have lengthened with diminished perceptual efficiency caused by fatigue. This would cause the incorrectly classified short cognitive fixations to be correctly reclassified as cognitive fixations with the onset of fatigue. Operators more resistant to fatigue would maintain shorter cognitive fixations which would re-main incorrectly classified as express fixations. The result is that those who struggled less were less likely to have their incorrectly classified cognitive fixations (those accidentally considered express fixations) reclassified as cognitive fixations; the outcome indicated by the fixation duration binning metric changed qualitatively when data quality was poor. Thus, maintaining reasonable quality for this metric is essential.

Percent Determinism (RQA)

Due to difficulties with the quality of the eye tracking and the expert level of experience with this method required to mitigate some of the resulting difficulties in analysis, a decision was made to exclude this metric from further consideration. To undertaking training required to understand the proper steps to take with imperfect data for this analysis is beyond the current scope.

Dwell Time Based Reliance

Finally, the AOI-based eye metric is a measure of fatigue to the extent that reliance behavior changes with level of fatigue. Specifically, misuse or disuse of automation may result from passive fatigue, which could be observed as an increase or decrease in AOI. The current effort found no change with time, nor differences between automation reliability conditions. Further, there was no correlation between dwell time based reliance and task engagement, performance based accuracy, or even performance-based reliance. The lack of correlation with task engagement would be explained if participants chose disparate effort-minimizing strategies in response to fatigue. Some participants may have trusted the automation, increasing the relative time spent in the response window, whereas

others may have chosen not to expend effort accounting for the recommendation of the automation and instead decided simply to do each task manually. Performance based accuracy may not have been correlated because reliability condition was necessary collapsed across condition because an online metric would not, in practice, necessarily have access to level of automation reliability level. The lack of correlation of this eye tracking-based metric and the performance based metric may be due to differences in the way reliance was assessed for each.

The difficulty detecting temporal trends and lack of correlational relationships with other relevant variables might also be explained by poor edge tracking. The most troublesome area of eye tracking on the screen during calibration was consistently the left edge where the image task window was placed. Tracking was particularly poor in the top left corner where the image queue was situated and where images would appear once a queue item was selected. By contrast, the lower left corner usually had more successful tracking than the top two thirds of the left edge. This added variability from tracking quality discrepancy may have made it difficult to detect Image Analysis behaviors for some participants.

Impact of Automation Reliability

It was hypothesized that the low reliability automated decision aid would inspire less subjective trust than the high reliability automated aid, which was confirmed by a comparison of post task trust in automation between reliability conditions. It was also hypothesized that this lack of trust would be evident in different reliance behaviors toward each automation. For both the Image Analysis and Weapon Release Authorization tasks, a strong main effect for reliability condition indicated that participants relied considerably less on automation in the low reliability condition. Dwell time based reliance was inconclusive; however, it is worth noting that all means for the low reliability condition were lower than the corresponding means for the high reliability condition.

It was hypothesized that participants might need to allocate extra effort to tasking supported by unreliable automation. For example, participants might double check their initial answer if it disagreed with the recommendation of the automation, even if they know the automation to be unreliable. Therefore, it was expected that participants in the low reliability condition would experience higher workload as well as higher task engagement. Participants did report higher workload particularly for mental and temporal demand. There was no reported difference in task engagement between conditions; however, participants in the low reliability condition exhibited higher levels of worry following the task.

Eye tracking metrics for mean fixation duration and frequency of overlong fixations showed marginally significant effects for reliability. However, these outcomes are difficult to interpret. Participants in the low reliability condition exhibited longer fixations on average, which may indicate higher levels of processing. This point of view may be supported by the higher number of overlong fixations, which, in addition to loss of concentration or daydreaming, may be indicative of concentration on a critical signal (Velichkovsky et al., 2002). An example might illustrate this interpretation best: a participant completing the Weapon Release Authorization task must interpret highly degraded images without moving closer to the screen in order to maintain eye tracking. An unreliable automation may call into question participant's judgements more often than the high reliability automation assuming that the participant is somewhat competent at the task. As a result, the participant may check the signals again, maintaining high levels of focus to decide, for example, if he or she really saw a short barrel or if the image degradation only made it seem that way. The limited amount of time to provide a response may have made such prolonged fixations necessary as an emergency event in driving was most successfully avoided by participants who maintained long fixations around the event in Velichkovsky et al.'s (2002) investigation of simulated driving.

How Impact of Reliability Changes with Fatigue

Finally, we expected the influence on reliance of reliability and fatigue to interact. Specifically, in accordance with the expectations laid out by Parasuraman and Riley (1997) and subsequent work concerning the contexts of misuse and disuse of automation, we hypothesized that those using high reliability automation would increase reliance with time on task (fatigue), and that those with low reliability automation would rely on automation less with time on task. The results of this effort did not fully support this hypothesis. The imaging tasks had a significant interaction; however, the lack of any clear linear pattern made it difficult to extrapolate from the data, yielding decidedly little support for our hypothesis. An interaction that did yield some insight was found for the Weapon Release Authorization task. Here, reliance on automation at both levels of reliability decreased with time; however, the decrease in reliance was less pronounced in the high reliability condition. Although there was the predicted interaction, complacency was not observed. Instead, both conditions experienced declines in reliance.

The tendency to disuse automation can be explained by previous work on passive fatigue. Specifically, deciding whether or not to rely on automation requires effort, and participants may have become reluctant to apply effort as they became more fatigued. Because the automation was imperfect (despite being quite reliable in the high reliability condition), participants may have chosen to simply complete the task themselves. Alternatively, the monotony of the task may have motivated participants to perform the tasks themselves to counter their boredom regardless of the reliability of the automation. In either case, the practical application is that fatigue may be especially associated with suboptimal use of automation when it would be best for the operator to be relying on it, i.e., when automation reliability is high. It was confirmed that reliance was generally positively correlated with performance accuracy in the high, but not the low, reliability condition. Interventions to improve reliance, therefore, should encourage operators to increase consideration of automated

suggestions. One possible intervention to encourage continued evaluation of the automation and facilitate its use might be to require operators to provide an evaluation of the automation recommendation with each decision event. For example, the operators might periodically indicate their confidence in the recommendation provided by the automation.

Concerning the influence of reliability on fatigue, it was hypothesized that the difference between reliability conditions in fatigue, as measured by eye tracking metrics, would increase with time. Those in the low reliability condition would remain engaged in tasking and therefore experience less fatigue relative to those who could rely on a reliable automated aid and therefor disengage. Only express fixations showed signs of an interaction, which was nonsignificant. Further, this interaction showed that instead of the predicted widening in level of fatigue between conditions, fatigue levels may have grown more similar with time. The higher, prolonged workload of the low reliability condition may have countered any greater engagement resulting from the need to remain personally engaged.

As a whole, these data suggest that it may be reasonable to expect similar reliance patterns with fatigue despite differences in automation reliability. These findings should be encouraging to those attempting to develop interventions to promote reliance optimization. Had reliance increased in the high reliability condition, a system would need to be able to gauge its own reliability in order to implement an appropriate intervention. Instead, it seems a common solution for low and high reliability automation may be effective.

Detection Algorithms

As a proof of concept, the final aim of this effort was to build algorithms, which could detect the onset of fatigue and the impact of fatigue on performance using only variables that could be used in actual mission environments. This precluded accuracy, event rate, and automation reliability. Thus, hierarchical regressions were conducted for the sixth, seventh, and eighth periods

using eye tracking metrics to predict subjective task engagement and performance on the Image Analysis and Weapon Release Authorization tasks. Three periods were used instead of the planned final period to circumvent the impact of a possible end spurt effect in the last period for which there is some evidence in the accuracy and fixation outcomes. The general idea was to use metrics that might be robust and easier to implement in the simple model, and in the complex model, add less proven and more calculation intensive metrics.

Generally the regressions were underpowered, yet there were some notable findings nonetheless. For predicting subjective fatigue, only the model for the eighth period was significant, which was driven by express fixation frequency. Notably, percentage of fixations falling within the cognitive range (150-900 ms) also contributed to this model though not significantly. Percent cognitive fixations did contribute to the sixth period and seventh period models, which had R^2 values of .25 and .21 but were not significant. Percent cognitive fixations also played a major role in the regressions predicting Image Analysis task accuracy with sizable contributions to two predictive complex models (sixth and seventh period). Interestingly, PERCLOS contributed to the seventh period model, as well as the nonsignificant eight period model. Dwell time-based reliance contributed to the sixth period model though not significantly. Express fixations again made a sizable contribution to the eighth period model, but for predicting accuracy this contribution was nonsignificant.

Taken together, percent cognitive fixations shows the most promise as a consistently predictive measure of both subjective fatigue and task performance. Express fixations also contributed especially to the eighth period model for both dependent variables. Other metrics contributed, but inconsistently. Dwell time-based reliance contributed to one period for each of the dependent variables but very little to other periods. PERCLOS drove prediction in the seventh and

eighth period models predicting performance, but had no prediction in the sixth model and moderate to weak effect sizes predicting task engagement.

Limitations

Eye tracking sample size and data quality were the biggest limitations in this effort. Despite some positive and informative outcomes, sizable effect sizes could not be verified. This study used sizable monitors, which tested well initially. Unfortunately, pilot testing may have benefitted from small sample sizes of lab members used to check for eye tracking issues, as edge tracking proved to be a reoccurring issue. Further, fatigue-based studies may be inherently difficult to conduct with an eye tracker do to mind-wandering and resulting gaze directions outside of the tasking area. Future work should reduce the size of the monitors to improve edge tracking and may consider adding instructions for participants to avoid avert their gazes from the interface. Another mitigating strategy would be to incorporate some kind of indicator of lost eye tracking in the display to encourage operators to reposition themselves for better eye tracking.

Another limitation in the current study is the variability in event frequency and automation reliability between periods. Although the intent was to simulate a realistic environment where some level of robustness was necessary for successful fatigue monitoring, a cleaner design would have been more theoretically useful as a first step considering the low rate of success with eye tracking. The impact of reliability can be observed in the correlations between accuracy and reliance (Table 3). The relationships between accuracy and reliance abruptly reversed in the sixth period for the image analysis task and in the seventh period of the Weapon Release Authorization task due to a drop in actual reliability for that period. For example, the automation's recommendation was only correct for one of the four Image Analysis tasks in the low reliability condition for period six (but 6/9 accurate in the previous period and 4/7 accurate in the following period).

Finally, the reliance metrics used in the current effort were not perfect reflection of how often a person utilizes the recommendation of the automation as might be assumed. The performance based reliance metric catches any sort of agreement due to factors that may be unrelated to reliance (Rice et al., 2010). For example, when a person is certain of a correct response, assuming that the certainty is well founded, the probability of agreement with automation will occur at the rate of reliability of the automation. Similarly, if the operator chooses not to heed the automation's recommendation in a situation of uncertainty, but instead decides to guess, there will be a certain amount of coincidental agreement as well. Although the agreement-based metric used in this effort was informative, a more discerning performance-based measure would better support the effort to evaluate an eye tracking-based measure.

The eye tracking-based reliance metric was also limited. The amount of time a participant gazed at the tasking section of the image analysis window reflected more than simply the time spent on tasking. For example, dwell time within the tasking area of the image analysis window included monitoring for queue population which might have waned with fatigue. High eye tracking quality may be required to create the smaller AOIs necessary to distinguish image gaze from queue gaze. Additionally, longer dwell time in the tasking area may not necessarily have indicated that reliance decreased. In other words, reliance rate may have remained constant and time taken to do each individual task may have increased with fatigue.

Future Work

The biggest practical contribution this effort was designed to make was the construction of informed algorithms for detecting operator fatigue in a multiple-task, UAV environment. Having identified promising metrics and demonstrated their ability to detect fatigue, future work can take the most promising metrics and apply them to adaptive intervention work. The outcomes of the

present study suggest that reliance may decrease in a low workload environment regardless of automation level, which can simplify the design of interventions in future work.

The limitations of this study also provide direction for future work. There is a need to understand how varying reliability (e.g., Ruff & Calhoun, 2011) impact eye tracking-based fatigue detection algorithms. This effort used multiple tasks as distractor items to generate a multiple task environment, but did not attempt to understand how attention is allocated (Cullen, Rogers, & Fisk, 2013) and other issues that may affect eye tracking based fatigue detection. Finally, future work could improve upon the method for detecting reliance objectively. A new metric might use a task that does not require extrapolation from agreement and accuracy to infer reliance. Such a metric might take advantage of other available information (e.g., accuracy, real time automation reliability, chance accuracy, subjective confidence) to distinguish between willful and incidental agreement with the automation recommendations.

CONCLUSION

Among many findings, two specific outcomes may be especially useful to future endeavors to support optimal reliance, which was the overarching goal of this effort. First, these findings suggested that it may be reasonable to expect similar reliance patterns with fatigue despite differences in automation reliability. These findings should be encouraging to those attempting to develop interventions to promote reliance optimization. Had reliance increased in the high reliability condition, a system would need to be able to gauge its own reliability in order to implement an appropriate intervention. Instead, it seems a common solution for low and high reliability automation may be effective. Second, the percentage of cognitive fixations eye tracking metric proved to be a predictive and, importantly, a robust measure. It drove prediction of regressions predicting task engagement and accuracy despite data quality and task consistency related challenges. Together, these key findings provide direction for future efforts to test and implement interventions informed by fatigue detection algorithms.

APPENDIX A: DEMOGRAPHICS QUESTIONNAIRE

Participant # _______ Date ___________ Age ______ Major ___________________
Gender ___

1. Do you have normal/corrected vision? YES NO
2. Are you in your usual state of health physically? YES NO
3. If NO, please briefly explain:
   ____
4. How many hours of sleep did you get last night? ______ hours
5. Have you had any caffeine in the last 12 hours? YES NO
6. What is your occupation? __________________________________
7. What is the highest level of education you have had?
   Less than 4 yrs of college ____ Completed 4 yrs of college ____ Other ____
8. When did you use computers in your education? (*Circle all that apply*)
   Grade School Jr. High High School
   Technical School College Did Not Use
9. Where do you currently use a computer? (*Circle all that apply*)
   Home Work Library Other________ Do Not Use
10. How many hours per day do you use a computer? ___________
11. Which of the following best describes your expertise with computers?
    Novice _____ Average_____ Proficient_____ Expert_____
12. How many hours per day do you use a computer? ___________
13. Estimate the **average number of hours per week** you have spent **playing all video games** within the **past two years (**e.g., PlayStation, Xbox, computer games)

0-1	2-4	5-7	8-10	11-13	14-16	17-19	20+
☐	☐	☐	☐	☐	☐	☐	☐

14. Estimate your **level of expertise** playing video games, in general
    (*0 = no expertise, 1=novice, 3 = intermediate, 6 = expert*)

0	1	2	3	4	5	6
☐	☐	☐	☐	☐	☐	☐

15. Estimate **average number of hours per week** you have spent **playing 'First Person Shooter' video games** within the **past two years (**e.g., *Call of Duty*)

0-1	2-4	5-7	8-10	11-13	14-16	17-19	20+
☐	☐	☐	☐	☐	☐	☐	☐

16. Estimate your **level of expertise** in playing First Person Shooter games

(*0 = no expertise, 1=novice, 3 = intermediate, 6 = expert*)

0	1	2	3	4	5	6
☐	☐	☐	☐	☐	☐	☐

17. Which First Person Shooter game have you played **the most**? (You may enter 'None')

____

18. Estimate **average number of hours per week** you have spent **playing other action video games** within the **past two years** (i.e, not First Person Shooter - e.g., *Grand Theft Auto*)

0-1	2-4	5-7	8-10	11-13	14-16	17-19	20+
☐	☐	☐	☐	☐	☐	☐	☐

19. Estimate your **level of expertise** in playing other action video games

(*0 = no expertise, 1=novice, 3 = intermediate, 6 = expert*)

0	1	2	3	4	5	6
☐	☐	☐	☐	☐	☐	☐

20. Which action video game have you played **the most**? (You may enter 'None')

____

APPENDIX B: NASA TASK-LOAD INDEX

INSTRUCTIONS: TLX RATINGS

We are interested in evaluating the experiences you had during the task. In the most general sense, we are examining the “workload” you experienced. The factors that influence workload may come from the task itself, your feelings about your own performance, how much effort you put in, or the stress and frustration you felt. The workload contributed by different task elements may change as you get more familiar with a task, perform easier or harder versions of it, or move from one task to another.

The following set of six rating scales was developed for you to use in evaluating your experiences during different tasks. Please read the descriptions of the scales carefully. If you have a question about any of the scales in the table, please ask the experimenter about it. It is extremely important that they be clear to you. You may keep the descriptions with you for reference during the experiment.

After performing the task, you will be presented with six rating scales. You are asked to evaluate the task by marking each scale at the point which matches your experience. Each line has two endpoint descriptors that describe the scale. You can place a cross on the line anywhere between the two endpoints. Note that “Performance” goes from “good” on the left to “bad” on the right. This order has been confusing for some people.

Please consider your responses carefully in distinguishing among different task conditions and consider each scale individually.

RATING SCALE DEFINITIONS

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

MENTAL DEMAND

Low High

PHYSICAL DEMAND

Low High

TEMPORAL DEMAND

Low High

PERFORMANCE

Good Poor

EFFORT

Low High

FRUSTRATION

Low High

APPENDIX C: DSSQ-3 STATE QUESTIONNAIRE

Instructions. This questionnaire is concerned with your feelings and thoughts at the moment. Please answer **every** question, even if you find it difficult. Answer, as honestly as you can, what is true of **you**. Please do not choose a reply just because it seems like the 'right thing to say'. Your answers will be kept entirely confidential. Also, be sure to answer according to how you feel **AT THE MOMENT**. Don't just put down how you usually feel. You should try and work quite quickly: there is no need to think very hard about the answers. The first answer you think of is usually the best.

Date today.................... Time of day now.............

For each statement, circle an answer from 0 to 4, so as to indicate how accurately it describes your feelings **AT THE MOMENT**.

Definitely false = 0, Somewhat false = 1,
Neither true nor false = 2, Somewhat true = 3, Definitely true = 4

1. I feel concerned about the impression I am making.	0	1	2	3	4
2. I feel relaxed.	0	1	2	3	4
3. The content of the task will be dull.	0	1	2	3	4
4. I am thinking about how other people might judge my performance.	0	1	2	3	4
5. I am determined to succeed on the task.	0	1	2	3	4
6. I feel tense.	0	1	2	3	4
7. I am worried about what other people think of me.	0	1	2	3	4
8. I am thinking about how I would feel if I were told how I performed.	0	1	2	3	4
9. Generally, I feel in control of things.	0	1	2	3	4
10. I am reflecting about myself.	0	1	2	3	4
11. My attention will be directed towards the task.	0	1	2	3	4
12. I am thinking deeply about myself.	0	1	2	3	4
13. I feel energetic.	0	1	2	3	4
14. I am thinking about things that happened to me in the past	0	1	2	3	4
15. I am thinking about how other people might perform on this task	0	1	2	3	4
16. I am thinking about something that happened earlier today.	0	1	2	3	4
17. I expect that the task will be too difficult for me.	0	1	2	3	4
18. I will find it hard to keep my concentration on the task.	0	1	2	3	4
19. I am thinking about personal concerns and interests.	0	1	2	3	4
20. I feel confident about my performance.	0	1	2	3	4
21. I am examining my motives.	0	1	2	3	4
22. I can handle any difficulties I may encounter	0	1	2	3	4

23. I am thinking about how I have dealt with similar tasks in the past	0	1	2	3	4
24. I am reflecting on my reasons for doing the task	0	1	2	3	4
25. I am motivated to try hard at the task.	0	1	2	3	4
26. I am thinking about things important to me.	0	1	2	3	4
27. I feel uneasy.	0	1	2	3	4
28. I feel tired.	0	1	2	3	4
29. I feel that I cannot deal with the situation effectively.	0	1	2	3	4
30. I feel bored.	0	1	2	3	4

POST-TASK QUESTIONNAIRE

Instructions. This questionnaire is concerned with your feelings and thoughts while you were performing the task. Please answer **every** question, even if you find it difficult. Answer, as honestly as you can, what is true of **you**. Please do not choose a reply just because it seems like the 'right thing to say'. Your answers will be kept entirely confidential. Also, be sure to answer according to how you felt **WHILE PERFORMING THE TASK**. Don't just put down how you usually feel. You should try and work quite quickly: there is no need to think very hard about the answers. The first answer you think of is usually the best.

For each statement, circle an answer from 0 to 4, so as to indicate how accurately it describes your feelings **WHILE PERFORMING THE TASK**.

Definitely false = 0, Somewhat false = 1,
Neither true nor false = 2, Somewhat true = 3, Definitely true = 4

1. I felt concerned about the impression I am making.	0	1	2	3	4
2. I felt relaxed.	0	1	2	3	4
3. The content of the task was dull.	0	1	2	3	4
4. I thought about how other people might judge my performance	0	1	2	3	4
5. I was determined to succeed on the task.	0	1	2	3	4
6. I felt tense.	0	1	2	3	4
7. I was worried about what other people think of me.	0	1	2	3	4
8. I thought about how I would felt if I were told how I performed	0	1	2	3	4
9. Generally, I felt in control of things.	0	1	2	3	4
10. I reflected about myself.	0	1	2	3	4
11. My attention was directed towards the task.	0	1	2	3	4
12. I thought deeply about myself.	0	1	2	3	4
13. I felt energetic.	0	1	2	3	4
14. I thought about things that happened to me in the past	0	1	2	3	4
15. I thought about how other people might perform on this task.	0	1	2	3	4
16. I thought about something that happened earlier today.	0	1	2	3	4
17. I found the task was too difficult for me.	0	1	2	3	4
18. I found it hard to keep my concentration on the task.	0	1	2	3	4
19. I thought about personal concerns and interests.	0	1	2	3	4
20. I felt confident about my performance.	0	1	2	3	4
21. I examined my motives.	0	1	2	3	4

22. I felt like I could handle any difficulties I encountered	0	1	2	3	4
23. I thought about how I have dealt with similar tasks in the past	0	1	2	3	4
24. I reflected on my reasons for doing the task	0	1	2	3	4
25. I was motivated to try hard at the task.	0	1	2	3	4
26. I thought about things important to me.	0	1	2	3	4
27. I felt uneasy.	0	1	2	3	4
28. I felt tired.	0	1	2	3	4
29. I felt that I could not deal with the situation effectively.	0	1	2	3	4
30. I felt bored.	0	1	2	3	4

APPENDIX D: HUMAN-COMPUTER TRUST SCALE

For each statement, circle an answer from 0 to 4, so as to indicate how accurately it describes your feelings.
CONSIDER ONLY THE TRIAL YOU JUST COMPLETED!

Extremely disagree = 0, Somewhat disagree = 1,

Neither disagree nor agree = 2, Somewhat agree = 3, Extremely agree = 4

1. The automation responds the same way under the same conditions at different times.

0 1 2 3 4

2. If I am not sure about a decision, I have faith that the automation will provide the best solution.

0 1 2 3 4

3. The advice the automation produces is as good as that which a highly competent person could produce

0 1 2 3 4

4. I understand how the automation will assist me with a decision I have to make.

0 1 2 3 4

5. I can rely on the automation to function properly.

0 1 2 3 4

6. I believe advice from the automation even when I don't know for certain that it is correct.

0 1 2 3 4

7. I like using the automation for decision making.

0 1 2 3 4

8. Although I may not know exactly how the automation works, I know how to use it to make decisions.

0 1 2 3 4

9. Overall, I trust the automation.

0 1 2 3 4

APPENDIX E: IRB LETTER OF APPROVAL

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Approval of Human Research

From: **UCF Institutional Review Board #1**
FWA00000351, IRB00001138

To: **Gerald Matthews** and Co-PIs: **Lauren Reinerman, Rebecca Leis, Ryan Wohleber**

Date: **April 01, 2015**

Dear Researcher:

On 4/1/2015, the IRB approved the following human participant research until 03/31/2016 inclusive:

Type of Review:	IRB Continuing Review Application Form Expedited Review
Project Title:	Sustaining Performance in Simulation UAV Operation: Pilot Study
Investigator:	Gerald Matthews
IRB Number:	SBE-13-09562
Funding Agency:	AFOSR, University of Cincinnati
Grant Title:	
Research ID:	1055976

The scientific merit of the research was considered during the IRB review. The Continuing Review Application must be submitted 30days prior to the expiration date for studies that were previously expedited, and 60 days prior to the expiration date for research that was previously reviewed at a convened meeting. Do not make changes to the study (i.e., protocol, methodology, consent form, personnel, site, etc.) before obtaining IRB approval. A Modification Form **cannot** be used to extend the approval period of a study. All forms may be completed and submitted online at https://iris.research.ucf.edu .

If continuing review approval is not granted before the expiration date of 03/31/2016, approval of this research expires on that date. When you have completed your research, please submit a Study Closure request in iRIS so that IRB records will be accurate.

Use of the approved, stamped consent document(s) is required. The new form supersedes all previous versions, which are now invalid for further use. Only approved investigators (or other approved key study personnel) may solicit consent for research participation. Participants or their representatives must receive a copy of the consent form(s).

All data, including signed consent forms if applicable, must be retained and secured per protocol for a minimum of five years (six if HIPAA applies) past the completion of this research. Any links to the identification of participants should be maintained and secured per protocol. Additional requirements may be imposed by your funding agency, your department, or other entities. Access to data is limited to authorized individuals listed as key study personnel.

In the conduct of this research, you are responsible to follow the requirements of the Investigator Manual.

On behalf of Sophia Dziegielewski, Ph.D., L.C.S.W., UCF IRB Chair, this letter is signed by:

Joanne Muratori

Signature applied by Joanne Muratori on 04/03/2015 05:08:47 PM EDT

IRB manager

APPENDIX F: INFORMED CONSENT

***Sustaining Performance in Simulation UAV Operation: Pilot Study* Informed Consent**

Principal Investigator(s): *Gerald Matthews, Ph.D.*

Co-Investigator(s): *Lauren Reinerman-Jones, Ph.D.*

Sponsor: *Air Force Office of Sponsored Research*

Investigational Site(s): *Institute for Simulation and Training*
University of Central Florida
3100 Research Parkway
Orlando, FL 32826

Introduction: Researchers at the University of Central Florida (UCF) study many topics. To do this we need the help of people who agree to take part in a research study. You are being invited to take part in a research study which will include 200 people at UCF. You must be 18 years of age or older to participate.

The investigator conducting this research is Dr. Gerald Matthews from the University of Central Florida's Institute for Simulation and Training.

What you should know about a research study:

- Someone will explain this research study to you.
- A research study is something you volunteer for.
- Whether or not you take part is up to you.
- You should take part in this study only because you want to.
- You can choose not to take part in the research study.
- You can agree to take part now and later change your mind.
- Whatever you decide it will not be held against you.
- Feel free to ask all the questions you want before you decide.

University of Central Florida IRB
IRB NUMBER: SBE-13-09562
IRB APPROVAL DATE: 04/01/2015
IRB EXPIRATION DATE: 3/31/2016

Purpose of the research study: The purpose of this study is to examine the impact of operating unmanned aerial vehicles (UAV) on the operator's performance, workload and state of mind, using a simulation of UAV control.

What you will be asked to do in the study:

Your task is to operate several UAVs within a simulated environment, assisted by computer automation. You will be trained in the use of UAV simulation software. You will be asked to sit at a computer workstation that consists of two computer displays and controls (mouse and keyboard). You may be asked to do any or all of the following tasks related to UAV operation: assign tasks to UAVs, choose new routes for the UAVs, identify and count targets in images, answer situation awareness questions that appear on the screen, respond to vehicle status events, and respond to unexpected aircraft. Following training you will be asked to perform a simulated UAV mission, lasting 120 min. You will also be asked to complete background questionnaires including information on your computer and video game experience. You will also be asked to complete questionnaires on your mood and attitudes towards the task before and after performance. When the study is over, the research assistant will give you debriefing information about the study. You do not have to answer every question or complete every task. You will not lose any benefits if you skip questions or tasks. During the task, a desk-mounted eye tracker will monitor eye position and movement. Before the actual experiment, the eye tracker will need to be calibrated to your gaze.

Location: Institute for Simulation and Training, Partnership II, Room 338.

Time required: We expect that you will be in this research study for up to 4 hours.

Funding for this study: This research study is being paid for by AFOSR.

Risks: There is a small risk that people who take part will develop what is ordinarily referred to as simulator sickness. It occurs once in a while to people who are exposed to prolonged continuous testing in simulated environments. Symptoms consist of nausea and a feeling of being light- headed. The risk is minimized as a result of the short duration of each session in the simulator. If you experience any of the symptoms mentioned, please tell the researcher and remain seated until the symptoms disappear.

Compensation or payment: Participants may expect to spend 4 hours performing experimental tasks, for which they will receive cash payment of $10/hr. for the amount of time they participate. Maximum course credit will be 4 credits.

Confidentiality: We will limit your personal data collected in this study to people who have a need to review this information. Organizations that may inspect and copy your information include the IRB and other representatives of UCF. In addition, because this research is sponsored by the Department of Defense and the U.S. Army, the Army Human Research Protections Office is eligible to review the research records. Data will be secured in locked cabinets at the Institute for Simulation and Training (IST) and disposed of following IRB protocol, which includes the shredding of all documents and proper deletion of electronic information.

Study contact for questions about the study or to report a problem: If you have questions, concerns, or complaints, or think the research has hurt you, talk to Dr. Gerald Matthews at 407-8820119 or at gmatthew@ist.ucf.edu.

IRB contact about your rights in the study or to report a complaint: Research at the University of
Central Florida involving human participants is carried out under the oversight of the Institutional Review Board (UCF IRB). This research has been reviewed and approved by the IRB. For information about the rights of people who take part in research, please contact: Institutional Review Board, University of Central Florida, Office of Research & Commercialization, 12201 Research Parkway, Suite 501, Orlando, FL 32826-3246 or by telephone at (407) 823-2901. You may also talk to them for any of the following:

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.
- You want to talk to someone besides the research team.
- You want to get information or provide input about this research.

REFERENCES

Abe, T., Nonomura, T., Komada, Y., Asaoka, S., Sasai, T., Ueno, A., & Inoue, Y. (2011). Detecting deteriorated vigilance using percentage of eyelid closure time during behavioral maintenance of wakefulness tests. *International Journal of Psychophysiology*, *82*(3), 269–274. http://doi.org/10.1016/j.ijpsycho.2011.09.012

Adams, J. A. (2007). Unmanned vehicle situation awareness: A path forward. In *Human systems integration symposium* (pp. 31–89). Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.119.4867&rep=rep1&type=pdf

Ahsberg, E., Gamberale, F., & Gustafsson, K. (2000). Perceived fatigue after mental work: an experimental evaluation of a fatigue inventory. *Ergonomics*, *43*(2), 252–268. http://doi.org/10.1080/001401300184594

Backs, R. W., Navidzadeh, H. T., & Xu, X. (2000). Cardiorespiratory indices of mental workload during simulated air traffic control. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 44, pp. 89–92). SAGE Publications. Retrieved from http://pro.sagepub.com/content/44/13/89.short

Bailey, N. R., & Scerbo, M. W. (2007). Automation-induced complacency for monitoring highly reliable systems: the role of task complexity, system experience, and operator trust. *Theoretical Issues in Ergonomics Science*, *8*(4), 321–348. http://doi.org/10.1080/14639220500535301

Banks, S., Jackson, M. L., & Van Dongen, H. P. A. (2012). Neuroscience of sleep and circadian rhythms. In G. Matthews, P. A. Desmond, C. Neubauer, & P. A. Hancock (Eds.), *The handbook of operator fatigue* (pp. 173–184). Burlington, VT: Ashgate Pub. Company.

Barnes, M. J., & Matz, M. F. (1998). Crew simulations for unmanned aerial vehicle (UAV) applications: sustained effects, shift factors, interface issues, and crew size. *Proceedings of the*

Human Factors and Ergonomics Society Annual Meeting, *42*(1), 143–147. http://doi.org/10.1177/154193129804200132

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 1252–1265.

Baumeister, R. F., Vohs, K. D., & Tice, D. M. (2007). The strength model of self-control. *Current Directions in Psychological Science*, *16*(6), 351–355. http://doi.org/10.1111/j.1467-8721.2007.00534.x

Bergum, B. O., & Lehr, D. J. (1963). End spurt in vigilance. *Journal of Experimental Psychology*, *66*(4), 383.

Billings, C. E. (1991). Toward a human-centered aircraft automation philosophy. *The International Journal of Aviation Psychology*, *1*(4), 261–270. http://doi.org/10.1207/s15327108ijap0104_1

Billings, C. E., Lauber, J. K., Funkhouser, Lyman, E. G., & Huff, E. M. (1976). *NASA aviation safety reporting system* (Quarterly Report No. NASA-TM-X-3445, A-6743, QR-76-1) (p. 24). Moffett Field, CA, United States: NASA Ames Research Center. Retrieved from http://www.researchgate.net/profile/Edward_Huff/publication/23616660_NASA_aviation_safety_reporting_system/links/02e7e534474285ba8a000000.pdf

Bone, E., & Bolkcom, C. (2003). *Unmanned aerial vehicles: Background and issues for congress* (p. 52). Washington, DC: Congressional Research Service The Library of Congress. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA467807

Bridgeman, B. (1992). Conscious vs Unconscious Processes: The Case of Vision. *Theory & Psychology*, *2*(1), 73–88. http://doi.org/10.1177/0959354392021004

Briest, S., Karrer, K., & Schleicher, R. (2006). Driving without awareness: Examination of the phenomenon. *Vision in Vehicles XI. Amsterdam: Elsevier Science Publishers BV*. Retrieved from http://www.isti.tu-

berlin.de/fileadmin/fg41/download/Team/Robert_Schleicher/Briest_Karrer_Schleicher_Driving_without_Awareness.pdf

Broadbent, D. E. (1971). *Decision and stress*. London, New York: Academic P.

Broderick, J. E., Schwartz, J. E., Vikingstad, G., Pribbernow, M., Grossman, S., & Stone, A. A. (2008). The accuracy of pain and fatigue items across different reporting periods: *Pain*, *139*(1), 146–157. http://doi.org/10.1016/j.pain.2008.03.024

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology*, *42*(3), 361–377. http://doi.org/10.1016/0301-0511(95)05167-8

Brown, I. D. (1997). Prospects for technological countermeasures against driver fatigue. *Accident Analysis & Prevention*, *29*(4), 525–531. http://doi.org/10.1016/S0001-4575(97)00032-8

Calhoun, G. L., Draper, M. H., Miller, C., Ruff, H. A., Breeden, C., & Hamell, J. (2013). Adaptable automation interface for multi-unmanned aerial systems control: Preliminary usability evaluation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 26–30. http://doi.org/10.1177/1541931213571008

Calhoun, G. L., Ruff, H. A., Draper, M. H., & Wright, E. J. (2011). Automation-level Transference Effects in simulated multiple unmanned aerial vehicle control. *Journal of Cognitive Engineering and Decision Making*, *5*(1), 55–82. http://doi.org/10.1177/1555343411399069

Chappelle, W. L., McDonald, K., & King, R. E. (2010). *Psychological attributes critical to the performance of MQ-1 Predator and MQ-9 Reaper U.S. Air Force Sensor Operators* (Final Report No. ADA525910). Retrieved from http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA525910

Chappelle, W. L., McDonald, K., & McMillan, K. (2011). *Important and critical psychological attributes of USAF MQ-1 Predator and MQ-9 Reaper pilots according to subject matter experts*. DTIC Document.

Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA545552

Chappelle, W. L., Salinas, A., & McDonald, K. (2011). *Psychological health screening of remotely piloted aircraft (RPA) operators and supporting units* (No. ADA582856). Wright-Patterson Air Force Base, OH: USAF School of Aerospace Medicine Department of Neuropsychiatry. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA582856

Chen, J. Y. C., & Barnes, M. J. (2012). Supervisory control of multiple robots: effects of imperfect automation and individual differences. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(2), 157–174. http://doi.org/10.1177/0018720811435843

Chen, J. Y. C., & Terrence, P. I. (2009). Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment. *Ergonomics*, *52*(8), 907–920. http://doi.org/10.1080/00140130802680773

Christodoulou, C. (2005). The assessment and measurement of fatigue. In J. DeLuca (Ed.), *Fatigue as a window to the brain* (pp. 19–35). Cambridge, Mass: MIT Press.

Christodoulou, C. (2012). Approaches to the measurement of fatigue. In G. Matthews, P. A. Desmond, C. Neubauer, & P. A. Hancock (Eds.), *The handbook of operator fatigue* (pp. 125–138). Burlington, VT: Ashgate Pub. Company.

Coifman, B., McCord, M., Mishalani, R. G., & Redmill, K. (2004). Surface transportation surveillance from unmanned aerial vehicles. In *Proc. of the 83rd Annual Meeting of the Transportation Research Board*. Retrieved from http://www.ugpti.org/smartse/research/citations/downloads/Coifman-Surface_Transport_Surveillance_from_UAV-2004.pdf

Craig, A., & Tran, Y. (2012). The influence of fatigue on brain activity. In G. Matthews, P. A. Desmond, C. Neubauer, & P. A. Hancock (Eds.), *The handbook of operator fatigue* (pp. 185–196). Burlington, VT: Ashgate Pub. Company.

Cullen, R. H., Rogers, W. A., & Fisk, A. D. (2013). Human performance in a multiple-task environment: Effects of automation reliability on visual attention allocation. *Applied Ergonomics*, *44*(6), 962–968. http://doi.org/10.1016/j.apergo.2013.02.010

Cummings, M. L., Brzezinski, A. S., & Lee, J. D. (2007). The impact of intelligent aiding for multiple unmanned aerial vehicle schedule management. Retrieved from http://dspace.mit.edu/handle/1721.1/90287

Cummings, M. L., Clare, A., & Hart, C. (2010). The role of human-automation consensus in multiple unmanned vehicle scheduling. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(1), 17–27. http://doi.org/10.1177/0018720810368674

Cummings, M. L., Mastracchio, C., Thornburg, K. M., & Mkrtchyan, A. (2013). Boredom and distraction in multiple unmanned vehicle supervisory control. *Interacting with Computers*, *25*(1), 34–47.

Curry, R. E., & Ephrath, A. R. (1976). Monitoring and control of unreliable systems. In T. B. Sheridan & G. Johannsen (Eds.), *Monitoring behavior and supervisory control* (pp. 193–203). Boston, MA: Springer US. Retrieved from http://link.springer.com/10.1007/978-1-4684-2523-9_17

Dahm, W. J. A. (2010). *Technology horizons: A vision for Air Force science and technology during 2010-2030* (Final Report No. AF/ST-TR-10-01-PR). United States Air Force.

de Greef, T., Lafeber, H., van Oostendorp, H., & Lindenberg, J. (2009). Eye movement as indicators of mental workload to trigger adaptive automation. In *Foundations of augmented cognition.*

Neuroergonomics and operational neuroscience (pp. 219–228). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-02812-0_26

Della Rocco, P. S. (1999). *The role of shift work and fatigue in air traffic control operational errors and incidents*. DTIC Document. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA360730

Desmond, P. A., & Hancock, P. A. (2001). Active and passive fatigue states. In P. A. Hancock & P. A. Desmond (Eds.), *Stress, workload, and fatigue* (pp. 455–465). Mahwah, N.J: Lawrence Erlbaum Associates.

Desmond, P. A., & Hoyes, T. W. (1996). Workload variation, intrinsic risk and utility in a simulated air traffic control task: Evidence for compensatory effects. *Safety Science*, *22*(1-3), 87–101. http://doi.org/10.1016/0925-7535(96)00008-2

Desmond, P. A., & Matthews, G. (1997). Implications of task-induced fatigue effects for in-vehicle countermeasures to driver fatigue. *Accident Analysis & Prevention*, *29*(4), 515–523. http://doi.org/10.1016/S0001-4575(97)00031-6

Desmond, P. A., & Matthews, G. (2009). Individual differences in stress and fatigue in two field studies of driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*(4), 265–276. http://doi.org/10.1016/j.trf.2008.12.006

De Vries, J., Michielsen, H. J., & Van Heck, G. L. (2003). Assessment of fatigue among working people: a comparison of six questionnaires. *Occupational and Environmental Medicine*, *60*(S1), i10–i15. http://doi.org/10.1136/oem.60.suppl_1.i10

Dinges, D. F. (1989). The nature of sleepiness: Causes, contexts, and consequences. In A. J. Stunkard & A. Baum (Eds.), *Eating, sleeping, and sex*. Hillsdale, N.J: L. Erlbaum Associates.

Dinges, D. F., Mallis, M. M., Maislin, G., & Powell, IV, J. W. (1998). *Evaluation of techniques for ocular measurement as an index of fatigue and the basis for alertness management* (Monograph No. DOT HS

808 762). National Highway Traffic Safety Administration. Retrieved from http://ntl.bts.gov/lib/21000/21900/21955/PB99150237.pdf

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(3), 474–486.

Dixon, S. R., Wickens, C. D., & Chang, D. (2004). Unmanned aerial vehicle flight control: False alarms versus misses. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 48, pp. 152–156). SAGE Publications. Retrieved from http://pro.sagepub.com/content/48/1/152.short

Dixon, S. R., Wickens, C. D., & Chang, D. (2005). Mission control of multiple unmanned aerial vehicles: A workload analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *47*(3), 479–487.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(4), 564–572. http://doi.org/10.1518/001872007X215656

Donmez, B., Cummings, M. L., & Graham, H. D. (2009). Auditory decision aiding in supervisory control of multiple unmanned aerial vehicles. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *51*(5), 718–729. http://doi.org/10.1177/0018720809347106

Donmez, B., Nehme, C., & Cummings, M. L. (2010). Modeling workload impact in multiple unmanned vehicle supervisory control. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *40*(6), 1180–1190. http://doi.org/10.1109/TSMCA.2010.2046731

Drury, J. L., Richer, J., Rackliffe, N., & Goodrich, M. A. (2006). *Comparing situation awareness for two unmanned aerial vehicle human interface approaches* (No. ADA456256). Bedford, MA: Mitre Corp. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA456256

Drury, J. L., Riek, L., & Rackliffe, N. (2006). A decomposition of UAV-related situation awareness. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot interaction* (pp. 88–94). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1121258

Drury, J. L., & Scott, S. D. (2008). Awareness in unmanned aerial vehicle operations. *The International C2 Journal*, *2*(1), 1–10.

Dureman, B. E. I., & BodéN, C. (1972). Fatigue in Simulated Car Driving. *Ergonomics*, *15*(3), 299–308. http://doi.org/10.1080/00140137208924432

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. http://doi.org/10.1016/S1071-5819(03)00038-7

Eggemeier, F. T., Wilson, G. F., Kramer, A. F., & Damos, D. L. (1991). Workload assessment in multi-task environments. In D. L. Damos (Ed.), *Multiple-task performance* (pp. 207–216). London ; Washington, DC: Taylor & Francis.

Eggers, J. W., & Draper, M. H. (2006). Multi-UAV control for tactical reconnaissance and close air support missions: operator perspectives and design challenges. In *Proc. NATO RTO Human Factors and Medicine Symp. HFM-135. NATO TRO, Neuilly-sur-Siene, CEDEX, Biarritz, France*. Retrieved from http://public.vrac.iastate.edu/~charding/audio/Eggers%20Paper%2011-06.pdf

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). In *Aerospace and Electronics Conference, 1988. NAECON 1988., Proceedings of the IEEE 1988 National* (pp. 789–795). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=195097

Endsley, M. R. (1996). Automation and situation awareness. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: theory and applications* (pp. 163–181). Mahwah, N.J: Lawrence Erlbaum Associates.

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*(3), 462–492. http://doi.org/10.1080/001401399185595

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(2), 381–394. http://doi.org/10.1518/001872095779064555

Endsley, M. R., & Rodgers, M. D. (1997). *Distribution of attention, situation awareness, and workload in a passive air traffic control task: Implications for operational errors and automation*. DTIC Document. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA328997

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Fidopiastis, C. M., Drexler, J., Barber, D., Cosenzo, K., Barnes, M., Chen, J. Y. C., & Nicholson, D. (2009). Impact of automation and task load on unmanned system operator's eye movement patterns. In *Foundations of augmented cognition. Neuroergonomics and operational neuroscience* (pp. 229–238). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-02812-0_27

Fischer, B., & Weber, H. (1993). What neural pathways mediate express saccades? *Behavioral and Brain Sciences*, *16*, 553–610.

Flemisch, F. O., & Onken, R. (2000). Detecting usability problems with eye tracking in airborne battle management support. In *RTO MP-57*. Oslo, Norway.

Funke, M. E. (2011). *Neuroergonomic and stress dynamics associated with spatial uncertainty during vigilance task performance* (Doctoral Dissertation). University of Cincinnati, Cincinnati, Ohio. Retrieved from http://rave.ohiolink.edu/etdc/view?acc_num=ucin1313773072

Galesic, M., Tourangeau, R., Couper, M. P., & Conrad, F. G. (2008). Eye-tracking data: New insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, *72*(5), 892–913. http://doi.org/10.1093/poq/nfn059

Galley, N., & Schleicher, R. (2004). Subjective and optomotoric indicators of driver drowsiness. In *Proceedings of the 3rd International Conference on Traffic and Transportation Psychology* (pp. 5–9). Retrieved from http://www.tu-berlin.de/fileadmin/fg41/download/Team/Robert_Schleicher/Galley_Schleicher_ICTTP2004.pdf

Gander, P. (2001). Fatigue management in air traffic control: the New Zealand approach. *Transportation Research Part F: Traffic Psychology and Behaviour*, *4*(1), 49–62. http://doi.org/10.1016/S1369-8478(01)00013-4

Gertler, J. (2012). *U.S. unmanned aerial systems* (No. ADA566235). Washington, DC: Library of Congress, Congressional Research Service. Retrieved from http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA566235

Gevins, A. S., Bressler, S. L., Cutillo, B. A., Illes, J., Miller, J. C., Stern, J., & Jex, H. R. (1990). Effects of prolonged mental work on functional brain topography. *Electroencephalography and Clinical Neurophysiology*, *76*(4), 339–350. http://doi.org/10.1016/0013-4694(90)90035-I

Gilbert, S. J., & Burgess, P. W. (2008). Executive function. *Current Biology*, *18*(3), R110–R114. http://doi.org/10.1016/j.cub.2007.12.014

Guan, Z., & Cutrell, E. (2007). An eye tracking study of the effect of target rank on web search. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 417–420). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1240691

Guznov, S. Y. (2011). *Visual search training techniques in a UAV simulator environment: Pilots' performance, workload, and stress*. University of Cincinnati. Retrieved from http://gradworks.umi.com/34/75/3475148.html

Guznov, S. Y., Matthews, G., Funke, G., & Dukes, A. (2011). Use of the RoboFlag synthetic task environment to investigate workload and stress responses in UAV operation. *Behavior Research Methods*, *43*(3), 771–780. http://doi.org/10.3758/s13428-011-0085-9

Haddal, C. C., & Gertler, J. (2010). *Homeland security: Unmanned aerial vehicles and border surveillance*. Washington, DC: Congressional Research Service, Library of Congress. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA524297

Hale, H. B., Williams, E. W., Smith, B. N., & Melton, C. E. (1971). Excretion patterns of air traffic controllers. *Aerospace Medicine*, *42*(2), 127–138.

Hancock, P. A., & Chignell, M. H. (1988). Mental workload dynamics in adaptive interface design. *IEEE Transactions on Systems, Man, and Cybernetics*, *18*(4), 647–658. http://doi.org/10.1109/21.17382

Hancock, P. A., & Desmond, P. A. (Eds.). (2001). *Stress, workload, and fatigue*. Mahwah, N.J: Lawrence Erlbaum Associates.

Hanson, M. L., & Harper, K. A. (2000). An intelligent agent for supervisory control of teams of unihabited combat air vehicles (UCAVs). Presented at the Unmanned Systems 2000 Conference, Orlando, FL.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0166411508623869

Haulman, D. L. (2003). *US unmanned aerial vehicles in combat, 1991-2003* (No. ADA434033). Air Force Historical Research Agency, Maxwell AFB, AL: DTIC Document. Retrieved from http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA434033

Helmreich, R. L., Merritt, A. C., & Wilhelm, J. A. (1999). The evolution of crew resource management training in commercial aviation. *The International Journal of Aviation Psychology*, *9*(1), 19–32. http://doi.org/10.1207/s15327108ijap0901_2

Helton, W. S. (2004). Validation of a Short Stress State Questionnaire. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(11), 1238–1242. http://doi.org/10.1177/154193120404801107

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

Hilgard, E. R. (1980). The trilogy of mind: Cognition, affection, and conation. *Journal of the History of the Behavioral Sciences*, *16*(2), 107–117. http://doi.org/10.1002/1520-6696(198004)16:2<107::AID-JHBS2300160202>3.0.CO;2-Y

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*(1-3), 73–93. http://doi.org/10.1016/S0301-0511(96)05223-4

Hockey, G. R. J. (2012). Challenges in fatigue and performance research. In G. Matthews, P. A. Desmond, C. Neubauer, & P. A. Hancock (Eds.), *The handbook of operator fatigue* (pp. 45–60). Burlington, VT: Ashgate Pub. Company.

Hockey, G. R. J., Wastell, D. G., & Sauer, J. (1998). Effects of sleep deprivation and user interface on complex performance: a multilevel analysis of compensatory control. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *40*(2), 233–253.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2015). *Eye tracking: a comprehensive guide to methods and measures*. Oxford: Oxford University Press.

Horne, J. A., & Reyner, L. A. (1996). Counteracting driver sleepiness: Effects of napping, caffeine, and placebo. *Psychophysiology*, *33*(3), 306–309. http://doi.org/10.1111/j.1469-8986.1996.tb00428.x

Hurter, C., Serrurier, M., Alonso, R., Tabart, G., & Vinot, J.-L. (2010). An automatic generation of schematic maps to display flight routes for air traffic controllers: structure and color optimization. In *Proceedings of the International Conference on Advanced Visual Interfaces* (pp. 233–240). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1843034

Ikegawa, S., Shinohara, M., Fukunaga, T., Zbilut, J. P., & Webber Jr, C. L. (2000). Nonlinear time-course of lumbar muscle fatigue using recurrence quantifications. *Biological Cybernetics*, *82*(5), 373–382.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. http://doi.org/10.1207/S15327566IJCE0401_04

Johns, M. W., Tucker, A., Chapman, R., Crowley, K., & Michael, N. (2007). Monitoring eye and eyelid movements by infrared reflectance oculography to measure drowsiness in drivers.

Somnologie - Schlafforschung Und Schlafmedizin, *11*(4), 234–242. http://doi.org/10.1007/s11818-007-0311-y

Johnson, R., Leen, M., & Goldberg, D. (2007). *Testing adaptive levels of automation (ALOA) for UAV supervisory control*. DTIC Document. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA470315

Kaplan, S., & Berman, M. G. (2010). Directed attention as a common resource for executive functioning and self-regulation. *Perspectives on Psychological Science*, *5*(1), 43–57. http://doi.org/10.1177/1745691609356784

Karrer, K., Briest, S., Vöhringer-Kuhnt, T., Baumgarten, T., & Schleicher, R. (2005). Driving without awareness. In G. Underwood (Ed.), *Traffic and transport psychology: theory and application ; proceedings of the ICTTP 2004* (1st ed). Amsterdam: Elsevier. Retrieved from https://books.google.com/books?hl=en&lr=&id=jrEKoWfOJc0C&oi=fnd&pg=PA455&dq=driving+without+awareness&ots=qDPVN9TiOm&sig=C6CPUcbGmlzi-aYSANTGqtxO7QM#v=onepage&q=driving%20without%20awareness&f=false

Kidwell, B., Calhoun, G. L., Ruff, H. A., & Parasuraman, R. (2012). Adaptable and adaptive automation for supervisory control of multiple autonomous vehicles. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 428–432. http://doi.org/10.1177/1071181312561096

Kircher, A., Uddman, M., & Sandin, J. (2002). *Vehicle control and drowsiness*. Swedish National Road and Transport Research Institute Linköping, Sweden. Retrieved from http://62.119.60.72/sv/publikationer/pdf/fordonskontroll-och-trotthet.pdf

Kloos, H., & Van Orden, G. (2010). Voluntary behavior in cognitive and motor tasks. *Mind and Matter*, *8*(1), 19–43.

Knipling, R. F., & Wierwille, W. W. (1994). *Vehicle-based drowsy driver detection: Current status and future prospects*. Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.460.3619&rep=rep1&type=pdf

Körber, M., Cingel, A., Zimmermann, M., & Bengler, K. (2015). Vigilance decrement and passive fatigue caused by monotony in automated driving. *Procedia Manufacturing*, *3*, 2403–2409. http://doi.org/10.1016/j.promfg.2015.07.499

Kozak, K., Curry, R., Greenberg, J., Artz, B., Blommer, M., & Cathey, L. (2005). Leading indicators of drowsiness in simulated driving. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 49, pp. 1917–1921). SAGE Publications. Retrieved from http://pro.sagepub.com/content/49/22/1917.short

Kramer, A. F., & McCarley, J. S. (2003). Oculomotor behaviour as a reflection of attention and memory processes: Neural mechanisms and applications to human factors. *Theoretical Issues in Ergonomics Science*, *4*(1-2), 21–55. http://doi.org/10.1080/14639220210159744

Lai, J.-S., Cella, D., Choi, S., Junghaenel, D. U., Christodoulou, C., Gershon, R., & Stone, A. (2011). How item banks and their application can influence measurement practice in rehabilitation medicine: A PROMIS fatigue item bank example. *Archives of Physical Medicine and Rehabilitation*, *92*(10), S20–S27. http://doi.org/10.1016/j.apmr.2010.08.033

Lal, S. K., & Craig, A. (2001). Electroencephalography activity associated with driver fatigue: Implications for a fatigue countermeasure device. *Journal of Psychophysiology*, *15*(3), 183.

Langan-Fox, J., Sankey, M. J., & Canty, J. M. (2009). Human factors measurement for future air traffic control systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *51*(5), 595–637. http://doi.org/10.1177/0018720809355278

Lavine, R. A., Sibert, J. L., Gokturk, M., & Dickens, B. (2002). Eye-tracking measures and human performance in a vigilance task. *Aviation, Space, and Environmental Medicine*, *73*(4), 367–372.

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270. http://doi.org/10.1080/00140139208967392

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153–184. http://doi.org/10.1006/ijhc.1994.1007

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. http://doi.org/10.1518/hfes.46.1.50_30392

Lewis, M., Wang, H., Chien, S. Y., Velagapudi, P., Scerri, P., & Sycara, K. (2010). Choosing autonomy modes for multirobot search. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(2), 225–233. http://doi.org/10.1177/0018720810366859

Lin, J., Wohleber, R. W., Matthews, G., Chiu, C.-Y. P., Calhoun, G. L., Ruff, H. A., & Funke, G. J. (2015). Video game experience and gender as predictors of performance and stress during supervisory control of multiple unmanned aerial vehicles. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*(1), 746–750. http://doi.org/10.1177/1541931215591175

Liu, D., Wasson, R., & Vincenzi, D. A. (2009). Effects of system automation management strategies and multi-mission operator-to-vehicle ratio on operator performance in UAV systems. *Journal of Intelligent and Robotic Systems*, *54*(5), 795–810. http://doi.org/10.1007/s10846-008-9288-4

MacDonald, W., A., & Hoffman, E., R. (1980). Review of relationships between steering wheel reversal rate and driving task demand. *Human Factors*, *22*(6), 733–739. http://doi.org/10.1177/001872088002200609

Mackie, R. R., Parasuraman, R., & Davies, R. D. (Eds.). (1977). A taxonomic analysis of vigilance performance. In *Vigilance theory, operational performance, and physiological correlates* (pp. 559–574). Boston, MA: Springer US. Retrieved from http://dx.doi.org/10.1007/978-1-4684-2529-1

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(2), 241–256.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In *Proceedings of Eleventh Australasian Conference on Information Systems* (pp. 6–8). Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.3874&rep=rep1&type=pdf

Mascord, D. J., & Heath, R. A. (1992). Behavioral and physiological indices of fatigue in a visual tracking task. *Journal of Safety Research*, *23*(1), 19–25. http://doi.org/10.1016/0022-4375(92)90036-9

Matthews, G., Campbell, S. E., Falconer, S., Joyner, L. A., Huggins, J., Gilliland, K., … Warm, J. S. (2002). Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. *Emotion*, *2*(4), 315–340. http://doi.org/10.1037//1528-3542.2.4.315

Matthews, G., Davies, R. D., Westerman, S. J., & Stammers, R. B. (2000). *Human performance: cognition, stress, and individual differences*. Hove, East Sussex : Philadelphia, PA: Psychology Press ; Taylor& Francis Group.

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

Matthews, G., Desmond, P. A., & Hitchcock, E. M. (2012). Dimensional models of fatigue. In G. Matthews, P. A. Desmond, C. Neubauer, & P. A. Hancock (Eds.), *The handbook of operator fatigue* (pp. 139–154). Burlington, VT: Ashgate Pub. Company.

Matthews, G., Desmond, P. A., Neubauer, C., & Hancock, P. A. (2012). An overview of operator fatigue. In G. Matthews, P. A. Desmond, C. Neubauer, & P. A. Hancock (Eds.), *The handbook of operator fatigue* (pp. 1–23). Burlington, VT: Ashgate Pub. Company.

Matthews, G., Hancock, P. A., & Desmond, P. A. (2012). Models of individual differences in fatigue for performance research. In G. Matthews, P. A. Desmond, C. Neubauer, & P. A. Hancock (Eds.), *The handbook of operator fatigue* (pp. 155–170). Burlington, VT: Ashgate Pub. Company.

Matthews, G., Joyner, L. A., Gilliland, K., Campbell, S. E., Falconer, S., & Huggins, J. (1999). Validation of a comprehensive stress state questionnaire: Toward a state "Big Three"? In I. Mervielde, I. J. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 335–350). Tilburg, Netherlands: Tilburg University Press.

Matthews, G., Schwean, V. L., Campbell, S. E., Saklofske, D. H., & Mohamed, A. A. R. (2000). Personality, self-regulation, and adaptation. In M. Boekarts, P. R. Printrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 171–207). New York: Academic Press. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/B9780121098902500354

Matthews, G., Szalma, J. L., Panganiban, A. R., Neubauer, C., & Warm, J. S. (2013). Profiling task stress with the Dundee Stress State Questionnaire. In L. Cavalcanti & S. Azevedo (Eds.), *Psychology of stress: New research* (pp. 49–91). Hauppauge, N.Y: Nova Science Publishers.

Matthews, G., Warm, J. S., Reinerman-Jones, L. E., Langheim, L. K., Guznov, S., Shaw, T. H., & Finomore, V. S. (2011). The functional fidelity of individual differences research: the case for context-matching. *Theoretical Issues in Ergonomics Science*, *12*(5), 435–450. http://doi.org/10.1080/1463922X.2010.549247

Matthews, G., Warm, J. S., Reinerman-Jones, L. E., Langheim, L. K., Washburn, D. A., & Tripp, L. (2010). Task engagement, cerebral blood flow velocity, and diagnostic monitoring for sustained attention. *Journal of Experimental Psychology: Applied*, *16*(2), 187–203. http://doi.org/10.1037/a0019572

Matthews, G., Warm, J. S., Reinerman, L. E., Langheim, L. K., & Saxby, D. J. (2010). Task engagement, attention, and executive control. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook of individual differences in cognition* (pp. 205–230). New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-1-4419-1210-7_13

Mayer, J. D., Chabot, H. F., & Carlsmith. (1997). Conation, affect, and cognition in personality. In G. Matthews (Ed.), *Cognitive science perspectives on personality and emotion*. Amsterdam ; New York: Elsevier.

May, J. F., & Baldwin, C. L. (2009). Driver fatigue: The importance of identifying causal factors of fatigue when considering detection and countermeasure technologies. *Transportation Research Part F: Traffic Psychology and Behaviour*, *12*(3), 218–224. http://doi.org/10.1016/j.trf.2008.11.005

McCrae, R. R., & Costa, Jr., P. T. (2008). Empirical and theoretical status of the five-factor model of personality traits. In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment* (1st ed, Vol. 1, pp. 273–294). Los Angeles, CA: SAGE Publications.

McKinley, R. A., McIntire, L. K., Schmidt, R., Repperger, D. W., & Caldwell, J. A. (2011). Evaluation of eye metrics as a detector of fatigue. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(4), 403–414. http://doi.org/10.1177/0018720811411297

Melton, C. E. (1982). *Physiological stress in air traffic controllers: A review*. DTIC Document. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA123853

Merritt, S. M. (2011). Affective processes in human-automation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(4), 356–370. http://doi.org/10.1177/0018720811411912

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(2), 194–210. http://doi.org/10.1518/001872008X288574

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K. (2014). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 0018720814561675.

Metzger, U., & Parasuraman, R. (2001). The role of the air traffic controller in future air traffic management: An empirical study of active control versus passive monitoring. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(4), 519–528.

Michielsen, H. J., De Vries, J., & Van Heck, G. L. (2003). In search of personality and temperament predictors of chronic fatigue: a prospective study. *Personality and Individual Differences*, *35*(5), 1073–1087. http://doi.org/10.1016/S0191-8869(02)00319-7

Mikulka, P. J., Scerbo, M. W., & Freeman, F. G. (2002). Effects of a biocybernetic system on vigilance performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *44*(4), 654–664.

Miller, C. A., & Parasuraman, R. (2003). Beyond levels of automation: An architecture for more flexible human-automation collaboration. In *Proceedings of the Human Factors and Ergonomics*

Society Annual Meeting (Vol. 47, pp. 182–186). SAGE Publications. Retrieved from http://pro.sagepub.com/content/47/1/182.short

Miller, C. A., & Parasuraman, R. (2007). Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(1), 57–75.

Miyao, M., Hacisalihzade, S. S., Allen, J. S., & Stark, L. W. (1989). Effects of VDT resolution on visual fatigue and readability: an eye movement approach. *Ergonomics*, *32*(6), 603–614. http://doi.org/10.1080/00140138908966135

Morana, C., Ramdani, S., Perrey, S., & Varray, A. (2009). Recurrence quantification analysis of surface electromyographic signal: Sensitivity to potentiation and neuromuscular fatigue. *Journal of Neuroscience Methods*, *177*(1), 73–79. http://doi.org/10.1016/j.jneumeth.2008.09.023

Morris, T. L., & Miller, J. C. (1996). Electrooculographic and performance indices of fatigue during simulated flight. *Biological Psychology*, *42*(3), 343–360. http://doi.org/10.1016/0301-0511(95)05166-X

Mouloua, M., Gilson, R., & Hancock, P. A. (2003). Human-centered design of unmanned aerial vehicles. *Ergonomics in Design: The Quarterly of Human Factors Applications*, *11*(1), 6–11.

Mouloua, M., Gilson, R., Kring, J., & Hancock, P. A. (2001). Workload, situation awareness, and teaming issues for UAV/UCAV operations. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 45, pp. 162–165). SAGE Publications. Retrieved from http://pro.sagepub.com/content/45/2/162.short

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429–460. http://doi.org/10.1080/00140139608964474

Muscio, B. (1921). Is a fatigue test possible? *British Journal of Psychology. General Section*, *12*(1), 31–46. http://doi.org/10.1111/j.2044-8295.1921.tb00036.x

Myers, G. L., & Fisk, A. D. (1987). Training consistent task components: Application of automatic and controlled processing theory to industrial task training. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *29*(3), 255–268. http://doi.org/10.1177/001872088702900302

Naschitz, J. E., Sabo, E., Naschitz, S., Rosner, I., Rozenbaum, M., Priselac, R. M., … Yeshurun, D. (2002). Fractal analysis and recurrence quantification analysis of heart rate and pulse transit time for diagnosing chronic fatigue syndrome. *Clinical Autonomic Research*, *12*(4), 264–272. http://doi.org/10.1007/s10286-002-0044-8

Neubauer, C., Matthews, G., Langheim, L., & Saxby, D. (2012). Fatigue and voluntary utilization of automation in simulated driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(5), 734–746. http://doi.org/10.1177/0018720811423261

Neubauer, C., Matthews, G., Saxby, D., & Langheim, L. (2011). Individual differences and automation choice in simulated driving. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 1563–1567. http://doi.org/10.1177/1071181311551326

Nuthmann, A., Smith, T. J., Engbert, R., & Henderson, J. M. (2010). CRISP: A computational model of fixation durations in scene viewing. *Psychological Review*, *117*(2), 382–405. http://doi.org/10.1037/a0018924

O'Hanlon, J. F., & Kelley, G. R. (1977). Comparison of performance and physiological changes between drivers who perform well and poorly during prolonged vehicular operation. In R. R. Mackie (Ed.), *Vigilance* (pp. 87–109). Boston, MA: Springer US. Retrieved from http://link.springer.com/10.1007/978-1-4684-2529-1_6

Ouma, J. A., Chappelle, W. L., & Salinas, A. (2011). *Facets of occupational burnout among US Air Force active duty and National Guard/reserve MQ-1 predator and MQ-9 reaper operators*. DTIC Document. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA548103

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, *55*(9), 1059–1072. http://doi.org/10.1080/00140139.2012.691554

Palmer, S. E. (1999). *Vision science: photons to phenomenology*. Cambridge, Mass: MIT Press.

Panganiban, A. R. (2013). *Task load and evaluative stress in a multiple UAV control simulation: The protective effect of executive functioning ability*. University of Cincinnati. Retrieved from http://rave.ohiolink.edu/etdc/view?acc_num=ucin1378215257

Panganiban, A. R., & Matthews, G. (2014). Executive functioning protects against stress in UAV simulation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 994–998. http://doi.org/10.1177/1541931214581208

Parasuraman, R. (1976). Consistency of individual differences in human vigilance performance: An abilities classification analysis. *Journal of Applied Psychology*, *61*(4), 486.

Parasuraman, R. (1979). Memory load and event rate control sensitivity decrements in sustained attention. *Science*, *205*(4409), 924–927.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(3), 381–410. http://doi.org/10.1177/0018720810376055

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, *47*(4), 51–55.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." *The International Journal of Aviation Psychology*, *3*(1), 1–23. http://doi.org/10.1207/s15327108ijap0301_1

Parasuraman, R., & Mouloua, M. (Eds.). (1996). *Automation and human performance: theory and applications*. Mahwah, N.J: Lawrence Erlbaum Associates.

Parasuraman, R., Mouloua, M., & Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *38*(4), 665–679. http://doi.org/10.1518/001872096778827279

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253.

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

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 511–520. http://doi.org/10.1518/001872008X312198

Pavlas, D., Burke, C. S., Fiore, S. M., Salas, E., Jensen, R., & Fu, D. (2009). Enhancing unmanned aerial system training: A taxonomy of knowledge, skills, attitudes, and methods. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, *53*(26), 1903–1907. http://doi.org/10.1518/107118109X12524444083159

Poole, A., & Ball, L. J. (2006). Eye tracking in HCI and usability research. *Encyclopedia of Human Computer Interaction*, *1*, 211–219.

Pope, A. T., & Bogart, E. H. (1992). *Identification of hazardous awareness states in monitoring environments* (No. 921136). Warrendale, PA: SAE International. Retrieved from http://www.sae.org/technical/papers/921136

Pope, A. T., Bogart, E. H., & Bartolome, D. S. (1995). Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology*, *40*(1-2), 187–195. http://doi.org/10.1016/0301-0511(95)05116-3

Prinzel, L. J., Freeman, F. G., Scerbo, M. W., Mikulka, P. J., & Pope, A. T. (2003). Effects of a psychophysiological system for adaptive automation on performance, workload, and the event-related potential P300 component. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *45*(4), 601–614.

Rango, A., Laliberte, A., Steele, C., Herrick, J. E., Bestelmeyer, B., Schmugge, T., … Jenkins, V. (2006). Using unmanned aerial vehicles for rangelands: current applications and future potentials. *Environmental Practice*, *8*(03), 159–168.

Rice, S., & Geels, K. (2010). Using System-Wide Trust Theory to Make Predictions About Dependence on Four Diagnostic Aids. *The Journal of General Psychology*, *137*(4), 362–375. http://doi.org/10.1080/00221309.2010.499397

Rice, S., Keller, D., Trafimow, D., & Sandry, J. (2010). Retention of a Time Pressure Heuristic in a Target Identification Task. *The Journal of General Psychology*, *137*(3), 239–255. http://doi.org/10.1080/00221309.2010.484447

Riley, J. M., & Endsley, M. R. (2005). Situation awareness in HRI with collaborating remotely piloted vehicles. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 49, pp. 407–411). SAGE Publications. Retrieved from http://pro.sagepub.com/content/49/3/407.short

Riley, M. A., Balasubramaniam, R., & Turvey, M. T. (1999). Recurrence quantification analysis of postural fluctuations. *Gait & Posture*, *9*(1), 65–78.

Rosekind, M. R., Gander, P. H., Gregory, K. B., Smith, R. M., Miller, D. L., Oyung, R., … Johnson, J. M. (1996a). Managing fatigue in operational settings 1: Physiological considerations and countermeasures. *Behavioral Medicine*, *21*(4), 157–165. http://doi.org/10.1080/08964289.1996.9933753

Rosekind, M. R., Gander, P. H., Gregory, K. B., Smith, R. M., Miller, D. L., Oyung, R., … Johnson, J. M. (1996b). Managing fatigue in operational settings 2: An integrated approach. *Behavioral Medicine*, *21*(4), 166–170. http://doi.org/10.1080/08964289.1996.9933754

Rosekind, M. R., Gander, P. H., Gregory, K. B., Smith, R. M., Miller, D. L., Oyung, R., … Johnson, J. M. (1997). Managing fatigue in operational settings 2: An integrated approach. *Hospital Topics*, *75*(3), 31–35. http://doi.org/10.1080/00185868.1997.10543762

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

Ruff, H. A., & Calhoun, G. L. (2011). Impact of automation's level and reliability on multi-vehicle supervisory control task performance. In *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*. Retrieved from http://arc.aiaa.org/doi/pdf/10.2514/6.2011-1489

Ruff, H. A., Calhoun, G. L., Draper, M. H., Fontejon, J. V., & Guilfoos, B. J. (2004). *Exploring automation issues in supervisory control of multiple UAVs*. DTIC Document. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA436181

Ruff, H. A., Narayanan, S., & Draper, M. H. (2002). Human interaction with levels of automation and decision-aid fidelity in the supervisory control of multiple simulated unmanned air vehicles. *Presence: Teleoperators and Virtual Environments*, *11*(4), 335–351.

Russell, S. M. (2014). *Alternative indices of performance: An exploration of eye gaze metrics in a visual puzzle task* (Dissertation). Wright State University. Retrieved from https://etd.ohiolink.edu/!etd.send_file?accession=wright1401291701&disposition=inline

Saito, S. (1992). Does fatigue exist in a quantitative measurement of eye movements? *Ergonomics*, *35*(5-6), 607–615. http://doi.org/10.1080/00140139208967840

Sarason, I. G., Sarason, B. R., Keefe, D. E., Hayes, B. E., & Shearin, E. N. (1986). Cognitive interference: Situational determinants and traitlike characteristics. *Journal of Personality and Social Psychology*, *51*(1), 215–226. http://doi.org/10.1037/0022-3514.51.1.215

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. *Handbook of Human Factors and Ergonomics*, *2*, 1926–1943.

Sauer, J., Kao, C.-S., & Wastell, D. (2012). A comparison of adaptive and adaptable automation under different levels of environmental stress. *Ergonomics*, *55*(8), 840–853. http://doi.org/10.1080/00140139.2012.676673

Sauer, J., Wastell, D. G., Robert, G., Hockey, G. R. J., & Earle, F. (2003). Performance in a complex multiple-task environment during a laboratory-based simulation of occasional night work. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *45*(4), 657–670.

Saxby, D. J., Matthews, G., Hitchcock, E. M., & Warm, J. S. (2007). Development of active and passive fatigue manipulations using a driving simulator. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 51, pp. 1237–1241). Sage Publications. Retrieved from http://pro.sagepub.com/content/51/18/1237.short

Saxby, D. J., Matthews, G., Hitchcock, E. M., Warm, J. S., Funke, G. J., & Gantzer, T. (2008). Effect of active and passive fatigue on performance using a driving simulator. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 52, pp. 1751–1755). Sage Publications. Retrieved from http://pro.sagepub.com/content/52/21/1751.short

Saxby, D. J., Matthews, G., Warm, J. S., Hitchcock, E. M., & Neubauer, C. (2013). Active and passive fatigue in simulated driving: Discriminating styles of workload regulation and their safety impacts. *Journal of Experimental Psychology: Applied*, *19*(4), 287–300. http://doi.org/10.1037/a0034386

Schleicher, R., Galley, N., Briest, S., & Galley, L. (2008). Blinks and saccades as indicators of fatigue in sleepiness warnings: looking tired? *Ergonomics*, *51*(7), 982–1010. http://doi.org/10.1080/00140130701817062

Schmeichel, B. J., Vohs, K. D., & Baumeister, R. F. (2003). Intellectual performance and ego depletion: Role of the self in logical reasoning and other information processing. *Journal of Personality and Social Psychology*, *85*(1), 33–46. http://doi.org/10.1037/0022-3514.85.1.33

Schmidt, E. A., Schrauf, M., Simon, M., Fritzsche, M., Buchner, A., & Kincses, W. E. (2009). Drivers' misjudgement of vigilance state during prolonged monotonous daytime driving. *Accident Analysis & Prevention*, *41*(5), 1087–1093. http://doi.org/10.1016/j.aap.2009.06.007

Schulte, A., & Donath, D. (2011). Measuring self-adaptive UAV operators' load-shedding strategies under high workload. In *Engineering psychology and cognitive ergonomics* (pp. 342–351). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-21741-8_37

Shaw, T. H., Matthews, G., Warm, J. S., Finomore, V. S., Silverman, L., & Costa, P. T. (2010). Individual differences in vigilance: Personality, ability and states of stress. *Journal of Research in Personality*, *44*(3), 297–308. http://doi.org/10.1016/j.jrp.2010.02.007

Sheridan, T. B., & Parasuraman, R. (2005). Human-automation interaction. *Reviews of Human Factors and Ergonomics*, *1*(1), 89–129. http://doi.org/10.1518/155723405783703082

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. DTIC Document. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA057655

Siegmund, G. P., King, D. J., & Mumford, D. K. (1996). Correlation of steering behavior with heavy-truck driver fatigue. http://doi.org/10.4271/961683

Silverman, B. (1992). Human-computer collaboration. *Human-Computer Interaction*, *7*(2), 165–196. http://doi.org/10.1207/s15327051hci0702_2

Singh, I. L., Molloy, R., & Parasuraman, R. (1993). Automation-induced "complacency": Development of the Complacency-Potential Rating Scale. *The International Journal of Aviation Psychology*, *3*(2), 111–122. http://doi.org/10.1207/s15327108ijap0302_2

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review*, *4*(2), 108–131.

Smith, P. J., McCoy, C. E., & Layton, C. (1997). Brittleness in the design of cooperative problem-solving systems: The effects on user performance. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, *27*(3), 360–371.

Smolensky, M. W. (1990). *The effect of work load history on operational errors in air traffic control simulation: The hysteresis effect - expectancy perserverance of short-term memory overload?* Texas Tech University. Retrieved from http://repositories.tdl.org/ttu-ir/handle/2346/18923

Sperandio, J.-C. (1978). The regulation of working methods as a function of work-load among air traffic controllers. *Ergonomics*, *21*(3), 195–202. http://doi.org/10.1080/00140137808931713

Stern, J. A., Boyer, D., & Schroeder, D. (1994). Blink rate: a possible measure of fatigue. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *36*(2), 285–297.

Stern, J. A., Boyer, D., Schroeder, D., Touchstone, M., & Stoliarov, N. (1994). *Blinks, saccades, and fixation pauses during vigilance task performance: 1. Time on task*. DTIC Document. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA290600

Stern, R. M., Ray, W. J., & Quigley, K. S. (2001). *Psychophysiological recording* (2nd ed). Oxford [England] ; New York: Oxford University Press.

Straussberger, S. (1996). *Monotony in air traffic control-contributing factors and mitigation strategies*. Eurocontrol Experimental Centre. Retrieved from https://eurocontrol.int/eec/gallery/content/public/document/eec/report/2006/019_Monotony_in_ATC.pdf

Stulberg, A. N. (2007). Managing the unmanned revolution in the US Air Force. *Orbis*, *51*(2), 251–265.

Sugarman, R. C., & Cozad, C. P. (1972). *Road tests of alertness variables* (Monograph) (p. 84). Alexandria, VA: National Technical Information Service.

Sussman, E. D., Sugarman, R. C., & Knight, J. R. (1971). *Use of simulation in a study investigating alertness during long-distance, low-event driving* (Component No. HS-011 999) (pp. 27–32). Washington.

Szalma, J. L., & Taylor, G. S. (2011). Individual differences in response to automation: The five factor model of personality. *Journal of Experimental Psychology: Applied*, *17*(2), 71–96. http://doi.org/10.1037/a0024170

Szalma, J. L., Warm, J. S., Matthews, G., Dember, W. N., Weiler, E. M., Meier, A., & Eggemeier, F. T. (2004). Effects of sensory modality and task duration on performance, workload, and stress in sustained attention. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(2), 219–233.

Takens, F. (1981). Detecting strange attractors in turbulence. In D. A. Rand & L.-S. Young (Eds.), *Dynamical systems and turbulence, Warwick 1980: proceedings of a symposium held at the University of Warwick 1979-80* (pp. 366–381). Berlin ; New York: Springer-Verlag.

Temple, J. G., Warm, J. S., Dember, W. N., Jones, K. S., LaGrange, C. M., & Matthews, G. (2000). The effects of signal salience and caffeine on performance, workload, and stress in an abbreviated vigilance task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *42*(2), 183–194.

Teo, G., & Szalma, J. L. (2011). The effects of task type and source complexity on vigilance performance, workload, and stress. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 1180–1184. http://doi.org/10.1177/1071181311551246

Thackray, R. I. (1980). *Boredom and monotony as a consequence of automation: A consideration of the evidence relating boredom and monotony to stress*. DTIC Document. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA085069

Thackray, R. I. (1981). The stress of boredom and monotony: a consideration of the evidence. *Psychosomatic Medicine*, *43*(2), 165–176.

Thayer, R. E. (1989). *The biopsychology of mood and arousal*. New York: Oxford University Press.

Thiffault, P., & Bergeron, J. (2003). Fatigue and individual differences in monotonous simulated driving. *Personality and Individual Differences*, *34*(1), 159–176.

Thorndike, E. L. (1914). *Educational psychology, Vol 3: Mental work and fatigue and individual differences and their causes*. New York: Teachers College. Retrieved from http://content.apa.org/books/13796-000

Tice, D. M., Baumeister, R. F., Shmueli, D., & Muraven, M. (2007). Restoring the self: Positive affect helps improve self-regulation following ego depletion. *Journal of Experimental Social Psychology*, *43*(3), 379–384. http://doi.org/10.1016/j.jesp.2006.05.007

Tvaryanas, A. P. (2006). Human systems integration in remotely piloted aircraft operations. *Aviation, Space, and Environmental Medicine*, *77*(12), 1278–1282.

Tvaryanas, A. P., Lopez, N., Hickey, P., DaLuz, C., Thompson, W. T., & Caldwell, J. L. (2006). *Effects of shift work and sustained operations: Operator performance in remotely piloted aircraft (OP-REPAIR)* (Final No. HSW-PE-BR-TR-2006-0001). Brooks City-Base. TX: United States Air Force 331st Human Systems Wing. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA443145

Tvaryanas, A. P., & Macpherson, G. D. (2009). Fatigue in pilots of remotely piloted aircraft before and after shift work adjustment. *Aviation, Space, and Environmental Medicine*, *80*(5), 454–461. http://doi.org/10.3357/ASEM.2455.2009

Tvaryanas, A. P., & Thompson, W. T. (2006). Fatigue in military aviation shift workers: survey results for selected occupational groups. *Aviation, Space, and Environmental Medicine*, *77*(11), 1166–1170.

Vanderhaegen, F., Crevits, I., Debernard, S., & Millot, P. (1994). Human-machine cooperation: Toward an activity regulation assistance for different air traffic control levels. *International Journal of Human-Computer Interaction*, *6*(1), 65–104. http://doi.org/10.1080/10447319409526084

Van Erp, J. B. F. (2000). *Controlling unmanned vehicles: The human factors solution* (No. ADP010325) (p. 13). TNO Soesterberg, Netherlands: Human Factors Research Institute.

Van Orden, K. F., Jung, T.-P., & Makeig, S. (2000). Combined eye activity measures accurately estimate changes in sustained visual task performance. *Biological Psychology*, *52*(3), 221–240.

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. In *Proceedings of the*

XXVII Conference of the Cognitive Science Society (pp. 2283–2288). Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.486.9238&rep=rep1&type=pdf

Velichkovsky, B. M., Rothert, A., Kopf, M., Dornhöfer, S. M., & Joos, M. (2002). Towards an express-diagnostics for level of processing and hazard perception. *Transportation Research Part F: Traffic Psychology and Behaviour*, *5*(2), 145–156.

Verwey, W. B., & Zaidel, D. M. (2000). Predicting drowsiness accidents from personal attributes, eye blinks and ongoing driving behaviour. *Personality and Individual Differences*, *28*(1), 123–142.

Vohs, K. D., Baumeister, R. F., & Ciarocco, N. J. (2005). Self-regulation and self-presentation: Regulatory resource depletion impairs impression management and effortful self-presentation depletes regulatory resources. *Journal of Personality and Social Psychology*, *88*(4), 632–657. http://doi.org/10.1037/0022-3514.88.4.632

Walters, B., French, J., & Barnes, M. J. (2000). Modeling the effects of crew size and crew fatigue on the control of tactical unmanned aerial vehicles (TUAVs). In *Simulation Conference, 2000. Proceedings. Winter* (Vol. 1, pp. 920–924). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=899893

Wang, Q., Yang, J., Ren, M., & Zheng, Y. (2006). Driver fatigue detection: a survey. In *Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on* (Vol. 2, pp. 8587–8591). Dalian: IEEE. http://doi.org/10.1109/WCICA.2006.1713656

Warm, J. S., & Dember, W. N. (1998). Tests of vigilance taxonomy. In R. R. Hoffman, M. F. Sherrick, & J. S. Warm (Eds.), *Viewing psychology as a whole: The integrative science of William N. Dember.* (pp. 87–112). Washington: American Psychological Association. Retrieved from http://content.apa.org/books/10290-004

Warm, J. S., Dember, W. N., & Hancock, P. A. (1996). Vigilance and workload in automated systems. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: theory and applications* (pp. 183–200). Mahwah, N.J: Lawrence Erlbaum Associates.

Warm, J. S., Matthews, G., & Finomore, V. S. (2008). Vigilance, workload, and stress. In P. A. Hancock & J. L. Szalma (Eds.), *Performance under stress* (pp. 115–141). Aldershot, England ; Burlington, VT: Ashgate Pub.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 433–441. http://doi.org/10.1518/001872008X312152

Watson, D., & Clark, L. A. (1984). Negative affectivity: The disposition to experience aversive emotional states. *Psychological Bulletin*, *96*(3), 465–490. http://doi.org/10.1037/0033-2909.96.3.465

Webber Jr, C. L., Schmidt, M. A., & Walsh, J. M. (1995). Influence of isometric loading on biceps EMG dynamics as assessed by linear and nonlinear tools. *Journal of Applied Physiology*, *78*(3), 814–822.

Webb, T. L., & Sheeran, P. (2003). Can implementation intentions help to overcome ego-depletion? *Journal of Experimental Social Psychology*, *39*(3), 279–286. http://doi.org/10.1016/S0022-1031(02)00527-9

Wells, A., & Matthews, G. (2015). *Attention and emotion: a clinical perspective* (Classic edition). Hove, East Sussex ; New York, NY: Psychology Press, Taylor & Francis Group.

Wertheim, A. H. (1978). Explaining highway hypnosis: Experimental evidence for the role of eye movements. *Accident Analysis & Prevention*, *10*(2), 111–129. http://doi.org/10.1016/0001-4575(78)90019-2

Wickens, C., Dixon, S., Goh, J., & Hammer, B. (2005). *Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis*. DTIC Document. Retrieved from http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA446167

Wiegmann, D. A. (2002). Agreeing with automated diagnostic aids: A study of users' concurrence strategies. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *44*(1), 44–50.

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science*, *2*(4), 352–367. http://doi.org/10.1080/14639220110110306

Wiener, E. L. (1985). *Cockpit automation: In need of a philosophy* (No. 851956). Warrendale, PA: SAE International. Retrieved from http://www.sae.org/technical/papers/851956

Wierwille, W. W., Wreggit, S. S., Kirn, C. L., Ellsworth, L. A., & Fairbanks, R. J. (1994). *Research on vehicle-based driver status/performance monitoring; development, validation, and refinement of algorithms for detection of drvier drowsiness* (No. HS-808 247 VPISU ISE 94-04).

Wilson, G. F., & Russell, C. A. (2003). Operator functional state classification using multiple psychophysiological features in an air traffic control task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *45*(3), 381–389. http://doi.org/10.1518/hfes.45.3.381.27252

Wohleber, R. W., Matthews, G., Funke, G. J., & Lin, J. (2016). Considerations in physiological metric selection for online detection of operator state: A case study. In *Proceedings of International Conference on Human-Computer Interaction.* Toronto, Canada: Springer-Verlag.

Wohleber, R. W., Matthews, G., Reinerman-Jones, L. E., Panganiban, A. R., & Scribner, D. (2015). Individual differences in resilience and affective response during simulated UAV operations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *59*.

Wylie, C. D., Shultz, T., Miller, J. C., Mitler, M. M., & Mackie, R. R. (1996). *Commercial motor vehicle driver fatigue and alertness study: Project report* (Monograph No. FHWA-MC-97-002) (p. 562). Washington, DC: Federal Highway Administration.