

2018

The Effect of Task Load, Automation Reliability, and Environment Complexity on UAV Supervisory Control Performance

Sarah M. Sherwood
Embry-Riddle Aeronautical University

Follow this and additional works at: <https://commons.erau.edu/edt>



Part of the [Ergonomics Commons](#), [Military Vehicles Commons](#), and the [Space Vehicles Commons](#)

Scholarly Commons Citation

Sherwood, Sarah M., "The Effect of Task Load, Automation Reliability, and Environment Complexity on UAV Supervisory Control Performance" (2018). *Doctoral Dissertations and Master's Theses*. 434.
<https://commons.erau.edu/edt/434>

This Dissertation - Open Access is brought to you for free and open access by Scholarly Commons. It has been accepted for inclusion in Doctoral Dissertations and Master's Theses by an authorized administrator of Scholarly Commons. For more information, please contact commons@erau.edu.

**THE EFFECT OF TASK LOAD, AUTOMATION RELIABILITY, AND ENVIRONMENT
COMPLEXITY ON UAV SUPERVISORY CONTROL PERFORMANCE**

A dissertation proposal
submitted in partial fulfillment of
the requirements for the degree of
Doctor of Human Factors
2018

by

SARAH M. SHERWOOD
M.S., Embry-Riddle Aeronautical University, 2014
B.A., University of South Florida, 2011

Dahai Liu, Ph.D.
Dissertation Chair

Beth Blickensderfer, Ph.D.
Bert Boquet, Ph.D.
Joseph Coyne, Ph.D.
Eric Vaden, M.S.
Dissertation Committee

THE EFFECT OF TASK LOAD, AUTOMATION RELIABILITY, AND ENVIRONMENT
COMPLEXITY ON UAV SUPERVISORY CONTROL PERFORMANCE

By

Sarah Marie Sherwood

This dissertation was prepared under the direction of the candidate's Dissertation Committee Co-chairs,
Dr. Dahai Liu and Dr. Elizabeth Blickensderfer, and has been approved by the members of the
Dissertation Committee. It was submitted to the College of Arts and Sciences and was accepted in partial
fulfillment of the requirements for the degree of

Doctor of Philosophy in Human Factors

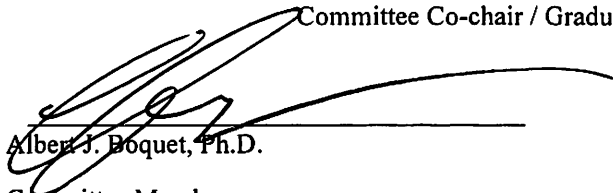


Dahai Liu, Ph.D.
Committee Co-chair



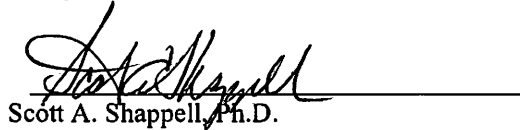
Elizabeth Blickensderfer, Ph.D.

Committee Co-chair / Graduate Program Coordinator



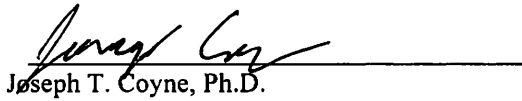
Albert J. Boquet, Ph.D.

Committee Member



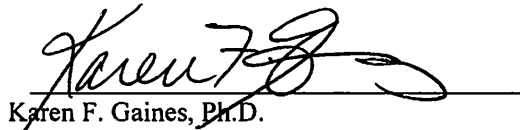
Scott A. Shappell, Ph.D.

Department Chair, Human Factors



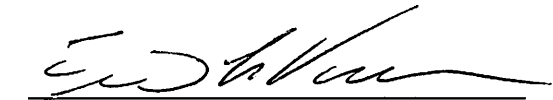
Joseph T. Coyne, Ph.D.

Committee Member



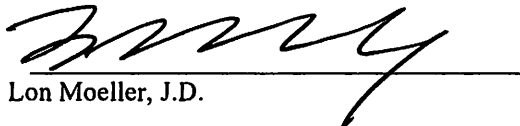
Karen F. Gaines, Ph.D.

Dean, College of Arts and Sciences



Eric A. Vaden, M.S.

Committee Member



Lon Moeller, J.D.

Senior Vice President for Academic Affairs and
Provost

11-26-18

Date

Acknowledgements

This dissertation was funded by the Naval Research Laboratory in support of the Intelligent Agent Support for Enterprise Decision Making Project, Contract No. GS35F009BA/N00173-14-F-0705. First and foremost, I would like to express my deepest appreciation for my committee chair, Dr. Dahai Liu, for his professional guidance, unwavering support, and frequent pep talks over tea. I would also like to extend my deepest gratitude to Dr. Joseph Coyne, Ms. Ciara Sibley, Mr. Jim Thomas, and Dr. Cyrus Foroughi at the Naval Research Lab for their financial support and instrumental assistance throughout the development and execution of this study; the criticality of their assistance with test bed coding, access to resources, data clean-up and analysis, and draft feedback to the success of this study cannot be understated. Of course, I am extremely grateful to the other members of my committee, Dr. Beth Blickensderfer, Dr. Bert Boquet, and Mr. Eric Vaden. Thank you all for your endless patience, insightful comments, and encouragement, and for generously offering your time, guidance, and support. I would also like to extend my sincere thanks to CDR Tatana Olsen, LT Mike Natali, Ms. Sabrina Drollinger, and Mr. Cory Moclaire at the Naval Aerospace Medical Institute (NAMI) for generously allowing me access to their lab space and participants, and for their support throughout the data collection process. I am also grateful to Dr. Kelly Neville and Dr. Jessica Cruit for their encouragement, late-night literature recommendations, and frequent reminders that the end is in sight. Finally, I would like to thank my family for their support and patience. It has been a long road, but I could not have traveled it without all of your support. Thank you all for seeing me to the end of this phase of my professional and personal journey and for molding me into a better researcher and a better person.

© Copyright by Sarah Sherwood 2018

All Rights Reserved

Abstract

Over the last decade, military unmanned aerial vehicles (UAVs) have experienced exponential growth and now comprise over 40% of military aircraft. However, since most military UAVs require multiple operators (usually an air vehicle operator, payload operator, and mission commander), the proliferation of UAVs has created a manpower burden within the U.S. military. Fortunately, simultaneous advances in UAV automation have enabled a switch from direct control to supervisory control; future UAV operators will no longer directly control a single UAV subsystem but, rather, will control multiple advanced, highly autonomous UAVs. However, research is needed to better understand operator performance in a complex UAV supervisory control environment. The Naval Research Lab (NRL) developed SCOUT™ (Supervisory Control Operations User Testbed) to realistically simulate the supervisory control tasks that a future UAV operator will likely perform in a dynamic, uncertain setting under highly variable time constraints. The study reported herein used SCOUT to assess the effects of task load, environment complexity, and automation reliability on UAV operator performance and automation dependence. The effects of automation reliability on participants' subjective trust ratings and the possible dissociation between task load and subjective workload ratings were also explored. Eighty-one Navy student pilots completed a 34:15 minute pre-scripted SCOUT scenario, during which they managed three helicopter UAVs. To meet mission goals, they decided how to best allocate the UAVs to locate targets while they maintained communications, updated UAV parameters, and monitored their sensor feeds and airspace. After completing training on SCOUT, participants were randomly sorted into low and high automation reliability groups. Within each group, task load (the number of messages and vehicle status updates that had to be made and the number of new targets that appeared) and environment complexity (the complexity of the payload monitoring task) were varied between low and high levels over the course of the scenario. Participants' throughput, accuracy, and expected value in response to mission events were used to assess their performance. In addition, participants

rated their subjective workload and fatigue using the Crew Status Survey. Finally, a four-item survey modeled after Lee and Moray's validated (1994) scale was used to assess participants' trust in the payload task automation and their self-confidence that they could have manually performed the payload task. This study contributed to the growing body of knowledge on operator performance within a UAV supervisory control setting. More specifically, it provided experimental evidence of the relationship between operator task load, task complexity, and automation reliability and their effects on operator performance, automation dependence, and operators' subjective experiences of workload and fatigue. It also explored the relationship between automation reliability and operators' subjective trust in said automation. The immediate goal of this research effort is to contribute to the development of a suite of domain-specific performance metrics to enable the development and/or testing and evaluation of future UAV ground control stations (GCS), particularly new work support tools and data visualizations. Long-term goals also include the potential augmentation of the current Aviation Selection Test Battery (ASTB) to better select future UAV operators and operational use of the metrics to determine mission-specific manpower requirements. In the far future, UAV-specific performance metrics could also contribute to the development of a dynamic task allocation algorithm for distributing control of UAVs amongst a group of operators.

Keywords: unmanned systems, supervisory control, automation, task load, complexity, automation reliability, operator performance, automation dependence, trust, workload, fatigue, reaction time, accuracy, throughput

Table of Contents

Acknowledgements.....	2
Copyright	3
Abstract.....	4
Purpose.....	9
2 Review of the Literature	17
2.1 Designing Automation for Performance	17
2.2 Automation Reliability and Signal Detection Theory	21
2.3 Trust in Automation	32
2.4 Environment Complexity.....	41
2.5 Task Load and Subjective Workload.....	44
2.6 Situation Awareness and UAV Ground Control Station Design	48
2.7 The Need for New Operator State and Performance Metrics	52
2.8 Supervisory Control Testing Environments	54
2.9 The Development of SCOUT	58
2.10 Literature Review Summary and Identified Gaps	62
3 Method.....	65
3.1 Participants.....	65
3.2 Apparatus	66
3.3 Procedure	67

UAV OPERATOR PERFORMANCE	7
3.4 Independent Variables	70
3.5 Dependent Variables	75
4 Results.....	78
4.1. Overview.....	78
4.2 UAV Operator Performance.....	80
4.3 UAV Operator Subjective Workload.....	95
4.4 UAV Operator Subjective Fatigue	103
4.5 UAV Operator Automation Dependence.....	105
4.6 UAV Operator Trust and Self-Confidence in Automation	110
5 Discussion	117
5.1 Overview.....	117
5.2 Effects of Task Load	117
5.3 Effects of Automation Reliability	120
5.4 Changes in Subjective Workload and Fatigue	123
5.5 Automation Dependence.....	125
5.6 Trust in Automation	127
6 Conclusion	131
7 Significance.....	136
References.....	138
Appendix A: Supervisory Control Operations User Testbed (SCOUT) Operation	151

A.1 Overview	151
A.2 Demographics and Initial Setup.....	152
A.3 Mission Training	155
A.4 Mission Components and Planning.....	156
A.4.1. UAV characteristics and capabilities.....	156
A.4.2. Target characteristics.....	158
A.4.3. Route planning.. ..	160
A.4.4. Route automation limitations.	163
A.4.5. Waypoints.....	163
A.4.6. Restricted operating zones (ROZs).....	164
A.5 Gameplay	167
A.5.1. Communication.....	167
A.5.2. Sensor orientation task.....	170
A.5.3. Reporting UAV position.....	171
A.5.4. Payload task.	173
A.5.5. Fatigue and workload questionnaires.....	175
A.6 Summary	177
Appendix B: Low-Cost Eye Tracking.....	178
B.1 Low-Cost Eye Tracking Systems	179

THE EFFECT OF TASK LOAD, AUTOMATION RELIABILITY, AND ENVIRONMENT COMPLEXITY ON UAV SUPERVISORY CONTROL PERFORMANCE

At present, multiple people are required to control most military unmanned aerial vehicles (UAVs); this has created a manpower burden that will be exacerbated as more unmanned systems come online. In 2005, UAVs represented only 5% of the U.S. military’s aircraft inventory. By 2010, that percentage rose to 41% (Gertler, 2012).

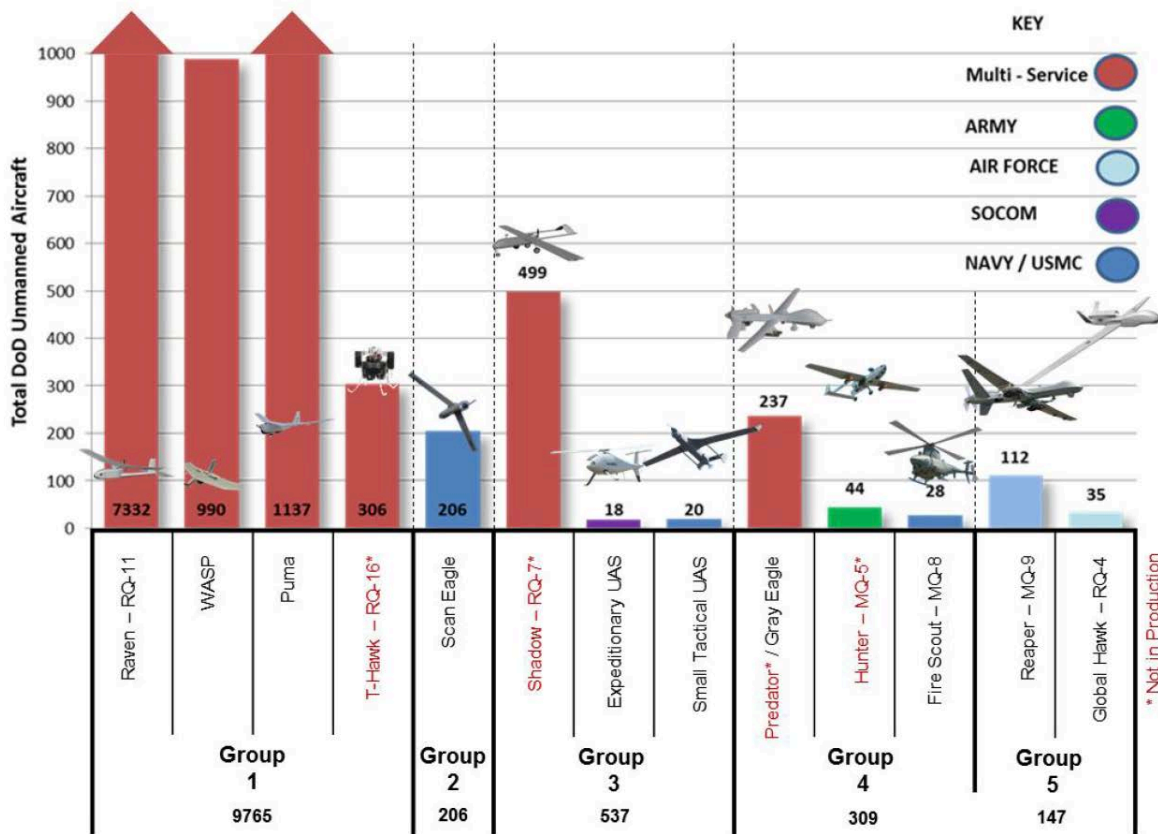


Figure 1.1. DoD UAV inventory as of July 1, 2013. Reprinted from the “Unmanned Systems Integrated Roadmap,” by the Department of Defense (2013). Note. Group 1 = micro/mini tactical UAVs, group 2 = small tactical UAVs, group 3 = tactical UAVs, group 4 = persistent UAVs, and group 5 = penetrating UAVs. Please see DoD (2013) page six for more information about these categories.

The DoD UAV inventory included 10,964 vehicles as of July 1, 2013 (Figure 1.1). This growth, along with parallel advances in UAV automation and interoperability, has catalyzed efforts to reduce the

manning requirements of UAV aircraft by moving from direct control of specific subsystems (e.g., payload and avionics) by multiple operators to single-operator supervisory control of multiple advanced, autonomous unmanned platforms. The shift toward supervisory control of unmanned systems is already underway; many operational unmanned systems feature waypoint navigation and mission management via maps (Sibley, Coyne, & Morrison, 2015).

Most current military unmanned platforms require three operators. The operators maintain one of three distinct roles: Mission Commander (MC), Air Vehicle Operator (AVO), and Payload Operator (PO). The mission commander is responsible for mission management, requesting access to controlled airspace, and communicating with customers of the services provided by the UAV. They are also responsible for disseminating information to the AVO and PO. The AVO is responsible for navigation (usually by managing waypoints) and monitoring the health and status of the vehicle. The PO is responsible for managing the platform's sensor suite and relaying relevant information to the MC and/or customer(s).

The task loads of the three roles are highly variable and, due to increased automation during certain mission phases leading to significant downtime for one or more of the roles, frequently unbalanced across the course of the mission. For example, the AVO might not interact with an unmanned platform at all while it is loitering over an area of interest because loitering is automated. In contrast, the PO is constantly tasked while the UAV is over target, as they are responsible for moving the sensor from one object of interest to another (Sibley, Coyne, & Morrison, 2015).

In addition to periods of unbalanced tasking, there are mission phases during which the entire crew is fully engaged. Close air support (CAS) missions, which are missions providing direct support to a ground unit engaged in combat, require significant human input and attention. The ground environment in a CAS mission is very fluid, with rapid and frequent changes in

friendly, enemy, and non-combatant locations. These entities cannot be identified exclusively with on-board sensors, so steady communication with the ground unit is often required. The ground commander's objectives can change rapidly in response to enemy actions or other changes in the ground environment, necessitating further communication with ground forces. Though efforts to pass some information via datalink are underway and some systems reviewed by the Naval Research Lab during the development of SCOUT heavily relied on chat messaging systems, communication is still conducted mostly via voice radio with personnel of varying experience levels and knowledge of the situation at hand. Airspace deconfliction is an additional source of task load in CAS missions, during which numerous aircraft are often operating within a small area. UAV operators must avoid overflight of friendly positions during weapons delivery and must deconflict from ground launched weapons and weapons delivered from other aircraft (Eggers & Draper, 2006; J. Coyne, personal communication, August 8, 2018).

Due to the high task load and dynamic, complex environment inherent to CAS missions, supervisory control of multiple UAVs by a single operator is not feasible for such missions. Mission management automation for CAS missions would require, at minimum, all of the aircraft, ground vehicles, personnel, weapons, tactical objectives, airspace, terrain, urban features, friendly, enemy, and non-combatants to be integrated in a digitized network. The creation of a networked data system sufficiently robust, detailed, and resistant to enemy exploitation to support a CAS mission is unlikely in the foreseeable future. The CAS environment benefits from the cognitive strengths and decision-making of human UAV operators, who are skilled at maintaining tactical SA and rapidly assessing and responding to ambiguous situations within the rules of engagement (Eggers & Draper, 2006).

On the other hand, Intelligence Surveillance and Reconnaissance (ISR) missions, the goal of which are to provide updated high-resolution imagery of predefined areas of interest, require minimal human input once the unmanned platform is airborne (Sibley, Coyne, & Sherwood, 2016). Most ISR mission targets are pre-planned, identified by accurate coordinates, and arrive at a manageable rate. Moreover, high-altitude ISR missions are often conducted at altitudes in excess of 60,000 feet, where most other aircraft cannot fly, few weather problems occur, and are beyond the range of many mobile threat systems. While the simulated ISR missions described herein feature rotary-winged UAVs operating at lower altitudes, the missions still take place in a relatively benign environment. While weather and surface-to-air threats become a concern at the altitudes at which rotary-winged aircraft operate, the simulated ISR missions, like most operational ISR missions, are still conducted within a well-defined environment subject to relatively few variations. Given the well-defined nature of the ISR mission environment, the essential mission tasks (e.g., navigation and aiming the sensor) are highly amenable to automation and ISR missions are thus an ideal use case for single-operator supervisory control of multiple unmanned systems (Eggers & Draper, 2006).

However, at present, military UAVs are still manned by a team of three operators (or more), regardless of mission type. The inefficiency and inflexibility of the current UAV control paradigm has influenced the DoD's desire to invert the ratio of operators to UAVs. Specifically, the 2015 Naval S&T Strategy calls for

The development of a distributed system of heterogeneous unmanned systems relying on network-centric, decentralized control that is flexible in its level of autonomy, with the ability to get the right level of information to the right echelon at the right time. (ONR, 2015, p. 28)

In the proposed decentralized network, groups of operators will share control of a large, distributed network of heterogeneous unmanned systems that are dynamically assigned based on theater mission requirements. Operators will no longer be statically assigned to a single task (e.g., payload operation) or vehicle, but will perform a common set of tasks for multiple heterogeneous platforms at different mission phases to accomplish mission objectives. This new paradigm is expected to increase both manning efficiency and flexibility (Sibley, Coyne, & Sherwood, 2016). For instance, flexible autonomy would allow for a hybrid approach of single and multiple aircraft control. The operator to vehicle ratio could be reduced for high task load combat missions, such as CAS, tactical reconnaissance (Tac Recce), Air Strike Control (ASC), and Combat Search and Rescue (CSAR). The most demanding of combat missions, such as CAS, might still require multiple operators to be assigned to each vehicle. Conversely, the operator to vehicle ratio could be increased for ISR missions, with a single operator assigned to multiple UAVs (Eggers & Draper, 2006).

Although ISR missions take place under *relatively* benign and predicable conditions, they are still nevertheless complex. In order to support the successful implementation of this new UAV management paradigm, research is needed to better understand operator performance during UAV supervisory control operations, especially under the variable time pressure (periods of both task overload and underload are common) and complexity intrinsic to the setting. Since the long-term goal of this research effort is the development of specialized UAV operator performance and status metrics to enable the development and evaluation of new work support tools (e.g., automated decision aids and data visualizations) and personnel selection, studies should be conducted within a testing environment that replicates the operational environment as closely as possible. The Naval Research Laboratory (NRL) developed SCOUT™

(Supervisory Control Operations User Testbed), a complex, realistic simulation environment, to fulfill this purpose (Sibley, Coyne, Avvari, Mishra, & Pattipati, 2016).

In the study described herein, SCOUT was used to evaluate the effects of variable task load, environment complexity, and automation reliability on operators' UAV supervisory control performance during a simulated ISR mission. More specifically, the study sought to support the following hypotheses:

- **Hypothesis 1:**
 - **H₀:** There is no significant effect of task load, payload task complexity, and automation reliability on operators' adjusted expected value on the UAV route-planning task. Expected value is defined as the sum of the products of each target's value and the probability of its successful location. Adjusted expected value compares this value for each participant against the performance of the participant with the best plan. See section 3.5.1.1 for more detailed information on the calculation of expected value.
 - **H_a:** There is a significant effect of task load, automation reliability, and payload task complexity on operators' adjusted expected value on the UAV route-planning task.
- **Hypothesis 2:**
 - **H₀:** There is no significant effect of task load, payload task complexity, and automation reliability on operators' accuracy on the payload task.
 - **H_a:** There is a significant effect of task load, payload task complexity, and automation reliability on operators' accuracy on the payload task.
- **Hypothesis 3:**
 - **H₀:** There is no significant effect of task load, payload task complexity, and automation reliability on operators' throughput on the communication task. Throughput is a composite

measure that accounts for both accuracy and response time. See section 3.5.1.2 for more information on throughput and its calculation.

- **H_a**: There is a significant effect of task load, payload task complexity, and automation reliability on operators' throughput on the communication task.
- **Hypothesis 4:**
 - **H_o**: There is no significant effect of task load, payload task complexity, and automation reliability on operators' subjective workload.
 - **H_a**: There is a significant effect of task load, payload task complexity, and automation reliability on operators' subjective workload.
- **Hypothesis 5:**
 - **H_o**: There is no significant effect of task load, payload task complexity, and automation reliability on operators' subjective fatigue.
 - **H_a**: There is a significant effect of task load, payload task complexity, and automation reliability on operators' subjective fatigue.
- **Hypothesis 6:**
 - **H_o**: There is no significant effect of task load, payload task complexity, and automation reliability on operators' automation dependence.
 - **H_a**: There is a significant effect of task load, payload task complexity, and automation reliability on operators' automation dependence.
- **Hypothesis 7:**
 - **H_o**: There is no significant effect of automation reliability on operator's subjective trust ratings of the automation.

- **H_a**: There is a significant effect of automation reliability on operator's subjective trust ratings of the automation.

2 Review of the Literature

2.1 Designing Automation for Performance

Table 2.1

Automation Functions and Human Cognition Equivalent

Automation Function Class	Level of Automation: Description	Four-Stage Model Equivalent	
Information acquisition	Low: mechanically moving sensors to scan and observe	Sensory processing	Positioning and orienting of sensory receptors, sensory processing, initial pre-processing of data prior to full perception, and selective attention
	Moderate: organization of incoming information/priority list with raw data visible/preserved		
	High: pre-filtered information brought to attention of operator (raw data not visible/preserved)		
Information analysis	Low: algorithms to extrapolate a variable over time (i.e., prediction)	Perception/working memory	Rehearsal, integration, and inference
	Medium: integration of multiple input variables into a single value		
	High: information manager that provides context-dependent summaries of data		
Decision and action selection	Selection from decision alternatives and implementation of decision (for ten levels of automation, see Table 2.2)	Decision making	Decision is made based on prior cognitive processing
Action implementation		Response selection	Implementation of response/action consistent with prior decision

Note. Adapted from “A Model for Types and Levels of Human Interaction with Automation,” by R. Parasuraman, T. B. Sheridan, and C. D. Wickens, 2000, *IEEE Transactions on Systems, Man, and Cybernetics- Part A: Systems and Humans*, 30(3), p. 287–289. Copyright 2000 by IEEE.

According to Parasuraman, Sheridan, and Wickens (2000), automation refers to “the full or partial replacement of a function previously carried out by the human operator” (p. 287). They proposed that automation could be applied to four classes of functions, which parallel the four-stage model of human information processing (Table 2.1). Automation can be applied to one

class of function (e.g., information acquisition), different combinations of functions (e.g., information acquisition and information analysis), or all four functions (Parasuraman, Sheridan, & Wickens, 2000).

Table 2.2

Sheridan and Verplank's 10 Levels of Automation of Decision and Action Selection

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

Note. Adapted from "A Model for Types and Levels of Human Interaction with Automation," by R. Parasuraman, T. B. Sheridan, and C. D. Wickens, 2000, *IEEE Transactions on Systems, Man, and Cybernetics- Part A: Systems and Humans*, 30(3), p. 287. Copyright 2000 by IEEE.

Within these functions, automation can be designed to supplant human activity to varying degrees. The role of automation in a human-machine system can best be described as a continuum of levels, ranging from full manual performance by a human operator to fully automated "black box" systems (Table 2.2). When designing automation for a specific system, the system designers must consider the following questions: (1) To what class of function(s) should the automation apply? (2) What level of automation should be applied within each function? (3) How reliable is the automation? And (4) what are the potential consequences of the automated decision/action(s)? (Parasuraman, Sheridan, & Wickens, 2000).

An important consideration for system designers in deciding on the function and level of automation in a human-machine system is the human performance consequences in the resulting

system. Particular types and levels of automation are evaluated by examining their human performance consequences, which include any potential impacts on: mental workload, situation awareness, complacency, and skill degradation. This requires a two-part process. First, initial research or modeling are needed to predict the upper and lower bounds of automation level; the upper bound is set at the level of automation at which human performance degrades relative to manual operation of the same task, and the lower bound is set at the point which automation at a given level is shown to improve human performance relative to manual operation of the task. Once the upper and lower bounds of automation level are determined for a specific system, the initial findings should be reevaluated while considering secondary criteria, most notably the reliability of the automation and the risk-level of the automated function (i.e., the severity of the consequences of the automated decision/action). Ideally, this process would continue in an iterative manner until the ideal automation level is determined, but real-world budget and timeline constraints could necessitate an abbreviation of the process (or even preclude it). Furthermore, other secondary factors that could come into play include ease of system integration, efficiency/safety trade-offs, liability issues, and the manufacturing and operating costs of the system itself (Parasuraman, Sheridan, & Wickens, 2000).

Considering that the entire purpose of an ISR mission is to locate and image targets of interest, system designers are much more likely to set a liberal response criterion for automated visual search aids because there is relatively low risk associated with a false alarm (see section 2.2 for more information on alert thresholds and signal detection theory). Unlike CAS missions, where a false alarm could result in erroneous weapons deployment (e.g., friendly fire or collateral damage), the risk associated with a false alarm on an ISR mission is relatively benign. From the perspective of system designers and other stakeholders, a miss on an ISR mission is

much more problematic because one or more missed critical targets can defeat the entire mission goal (e.g., Parasuraman, Sheridan, & Wickens, 2000). However, frequent false alarms are known to erode user trust in automation more so than misses (Bliss, 2003), and this erosion of trust can lead to the automation being underutilized with resulting increases in operator workload and decreases in operator performance.

It is therefore much more realistic, and useful, to search for indicators of operator performance and state that are sensitive enough to discriminate between the human performance impacts of automated aids with relatively small differences in reliability, as it is unlikely that highly unreliable automation will be considered by system designers fine-tuning the alert threshold during testing and evaluation of new search task automation for use on ISR missions.

Search aids designed for use in CAS missions, on the other hand, are much more likely to feature automation with a conservative response criterion because the consequences of a false alarm are potentially high. The Naval S&T Strategy (2015) calls for a distributed system of heterogeneous unmanned systems with *dynamically changing levels of autonomy*. As previously discussed, CAS missions require significant human input and attention; such missions presently require the full engagement of a three-operator team (the MC, the AVO, and the PO). Due to the high task load and dynamic, complex environment inherent to CAS missions, it is likely that such missions will require more direct operator control and will benefit from the resilience of human decision-making under conditions of uncertainty. Thus, for the foreseeable future, the manning requirement of UAVs engaged in CAS missions will remain high and automation appropriate for use under such conditions will likely fulfill information acquisition and analysis functions and will feature a relatively low LOA with preservation of raw data (Eggers & Draper, 2006; Parasuraman, Sheridan, & Wickens, 2000).

2.2 Automation Reliability and Signal Detection Theory

Automation reliability can be defined in terms of signal detection theory (SDT; Green & Swets, 1966; Zuniga, McCurry, & Trafton, 2014). SDT applies whenever a human operator or an automated system must discriminate between two possible states; in many cases, this task takes the form of discrimination between a signal (stimulus present) and noise (stimulus absent). For a given experimental trial, the human or computer decision-maker decides whether a signal is present or absent based on the value of a *decision variable*. If the decision variable is sufficiently high, a point defined by the value of the *criterion*, the decision-maker will indicate that a signal is present. Conversely, if the decision variable does not exceed the criterion, the decision-maker will indicate that a signal is absent (Nevin, 1969).

The distribution of the decision variable across trials is the *signal distribution*, whereas the distribution for noise trials is the *noise distribution*. However, since noise is always present, the signal distribution could more accurately be described as the *signal + noise distribution*. The regions of the signal and noise distributions that fall above the criterion value are the hit rate and false alarm rate, respectively. The regions of the signal and noise distributions that fall below the criterion value are the miss and correct rejection rates, respectively.

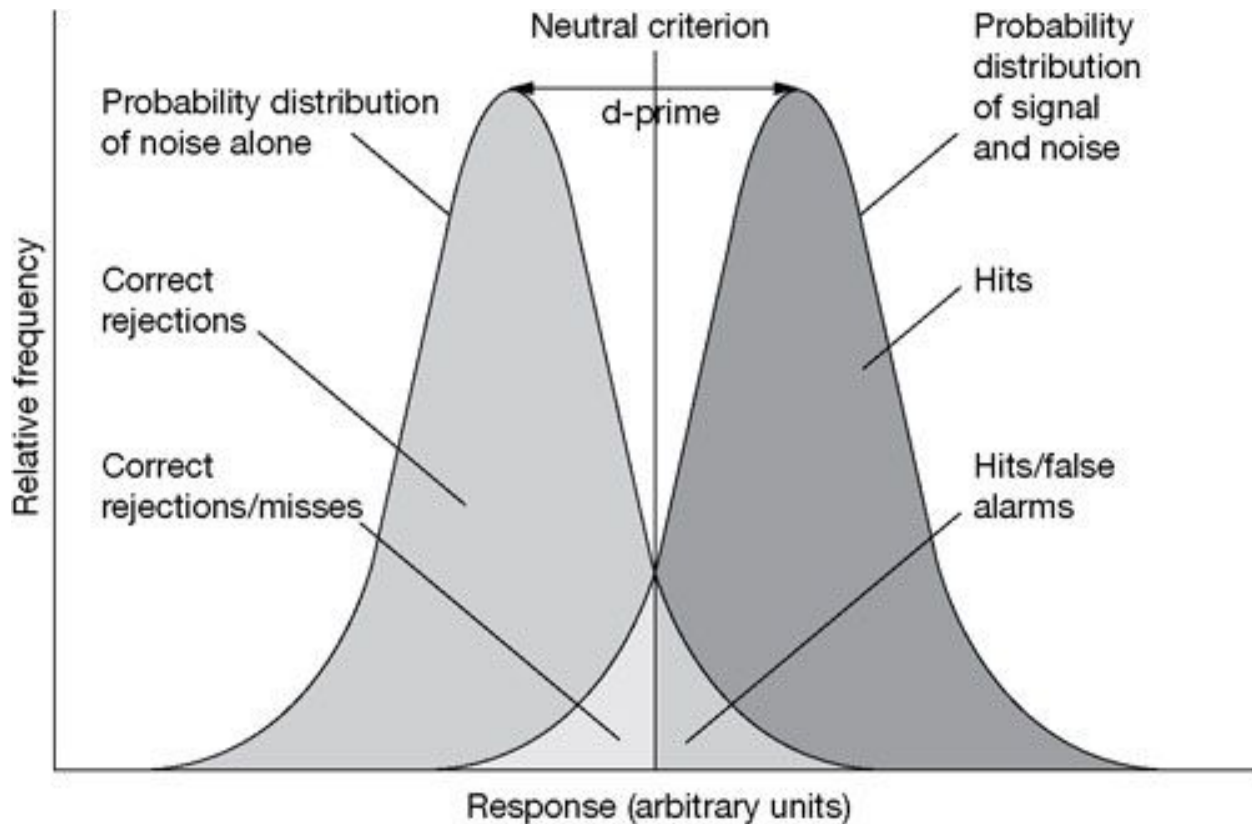


Figure 2.1. Noise distribution (left), signal + noise distribution (right), and d' -prime. Reprinted from “Noisy Patients’—Can Signal Detection Theory Help?” by R. Oliver, O. Bjoertomt, R. Greenwood, and J. Rothwell, 2008, *Nature Reviews Neurology*, 4(6), p. 306. Copyright 2008 by Springer Nature Ltd.

The discriminability of a signal, or d' (d-prime), is determined by both the degree of separation and the spread of the noise and signal distributions (Figure 2.1). More specifically, d' is the *sensitivity index* statistic used in signal detection theory. Mathematically, d' is defined as the separation between the signal and noise distributions divided by the spread of the distributions. The degree of separation will be larger for strong signals. The spread will be greater when there is less noise. A large spread and a large degree of separation will both result in less overlap between the signal and noise distributions. The degree of overlap is an inverse measure of sensitivity. A higher d' indicates a smaller overlap between the signal and noise distributions, a greater sensitivity and, thus, an easier discrimination task (Macmillan, 2002).

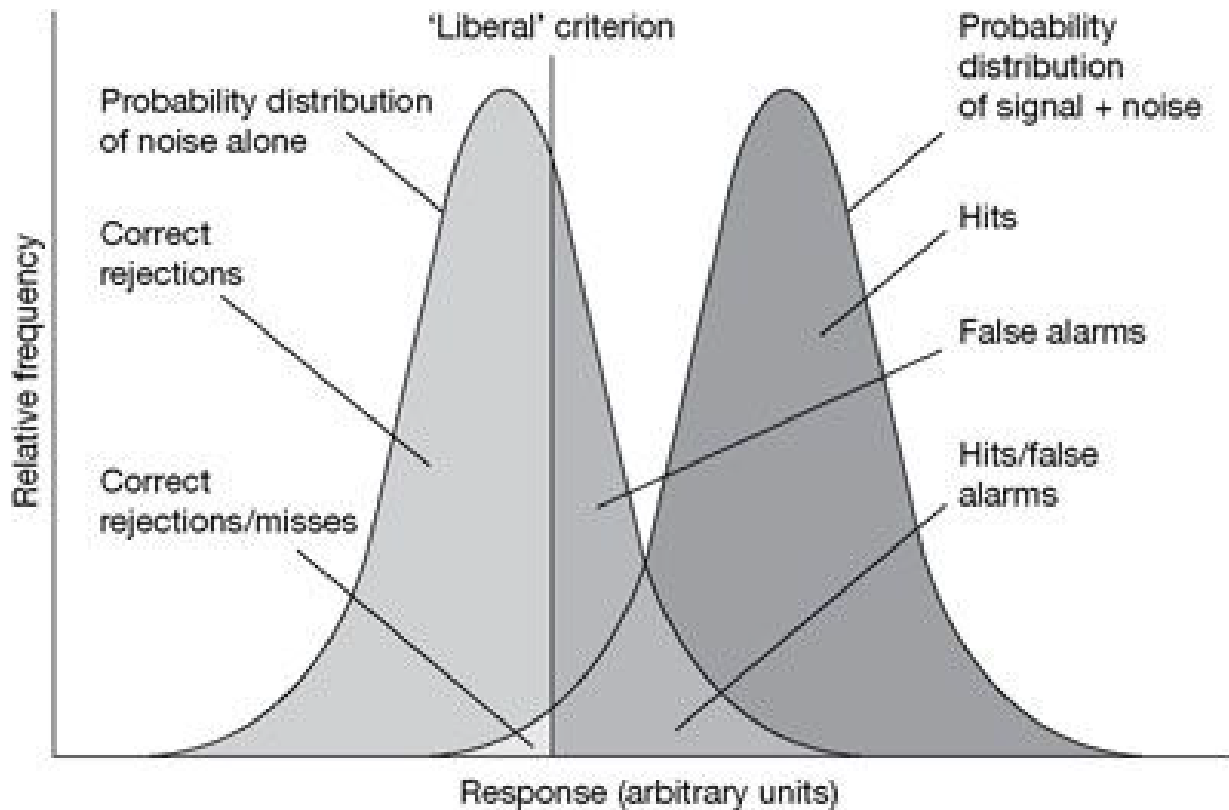


Figure 2.2. Hit, miss, false alarm, and correct rejection regions given a liberal response criterion. Reprinted from “Noisy Patients’—Can Signal Detection Theory Help?” by R. Oliver, O. Bjoertomt, R. Greenwood, and J. Rothwell, 2008, *Nature Reviews Neurology*, 4(6), p. 306. Copyright 2008 by Springer Nature Ltd.

However, sensitivity is not the only factor to influence hit and false alarm rates. Hit and false alarm rates are influenced by two factors: sensitivity (the degree of overlap between the signal and noise distributions) and the location of the criterion (c). The location of the criterion changes with the decision-maker’s tendency to indicate either the presence or absence of a signal, or their *response bias*. Response bias is independent of sensitivity. Negative values of c indicate a liberal response bias, or a tendency to indicate that a signal is present (Figure 2.2). One is more likely to observe a liberal response bias when the costs of a miss are higher than the costs of a false alarm, which is likely the case for ISR missions. A positive c value indicates a conservative response bias, or tendency to indicate that a target is absent. Conservative response

biases are more commonly observed when the cost of a false alarm is higher than the cost of a miss. For example, any target-detection prior to live weapons fire in a CAS mission would likely involve a very conservative response criterion. However, the response criterion for a target detection task in an ISR mission, where the consequences of a false alarm are less severe but a missed target would be more problematic, is likely to be more liberal and prone to false alarms.

Table 2.3

Automation Reliability in Terms of Signal Detection Theory (SDT)

AUTOMATION RELIABILITY	SIGNAL PRESENT	AUTOMATION RESPONSE	SDT OUTCOME
CORRECT FUNCTIONING (MORE RELIABLE)	X	X	Hit
			Correct Rejection
ERROR (LESS RELIABLE)	X		Miss
		X	False Alarm

Note. Noise is always present, regardless of whether the signal itself is present or absent.

As previously stated, automation reliability can be defined in terms of signal detection theory (SDT; Green & Swets, 1966; Zuniga, McCurry, & Trafton, 2014). Correct identification of a signal (e.g., a potential target) by the automation can be thought of as a hit, no response to noise in absence of a signal as a correct rejection, failure to respond to a signal as a miss, and erroneous response to noise in absence of a signal as a false alarm. Hits and correct rejections indicate correct automation functioning, while misses and false alarms indicate automation errors (Table 2.3). In the context of an automated system, sensitivity (d') is a measure of the combined rate of hits and false alarms of the system. Sensitivity indicates the accuracy of the system in differentiating the signal from the noise. However, systems of equal sensitivity can exhibit very different response patterns depending on the response criterion (Zuniga, McCurry, & Trafton, 2014).

In the study described herein, automation reliability is operationalized using SDT or, more specifically, as the percentage of correct responses (hits and correct rejections) for all trials. The automation employed in the high reliability condition has a 97.0% hit rate and 97.0% correct

rejection rate, which works out to an overall reliability of 97.0%. The automation in the low reliability condition possesses a more liberal response criterion; it has a 100.0% hit rate and 85.0% correct rejection rate, which works out to an overall reliability of 92.5%. This is problematic for participants, who are penalized for false alarms. SCOUT's sensor feed automation is generally subject to a more liberal response criterion (i.e., it is prone to a greater number of false alarms) since it is designed to simulate payload automation designed for an ISR mission.

There has been a substantial amount of research conducted on the impact of automation reliability on operator performance. However, most of the current research has been focused on the reliability of information acquisition and information analysis automation (e.g., Chancey, Bliss, Yamani, & Handley, 2016; Parasuraman, Molloy, & Singh, 1993; Rice, 2009; Rovira, McGarry, & Parasuraman, 2007; Wickens & Dixon, 2007; Wickens, Dixon, Goh, & Hammer, 2005; Dixon, Wickens, & McCarley, 2007) but relatively little attention has been paid to decision and action selection or action implementation automation (e.g., Calhoun et al., 2016; Ruff, Narayanan, & Draper, 2002).

Generally, research has shown that operator performance is increased in systems employing diagnostic automation when said automation is 80% or more reliable and operator performance is largely unaffected by automation with a 70% to 80% reliability. However, automated aids less than 70% reliable begin to negatively impact operator performance (Dixon & Wickens, 2006; Hillesheim & Rusnock, 2016; Maltz & Shinar, 2003; Parasuraman & Manzey, 2010; and Wickens & Dixon, 2007).

According to Meyer (2001, 2004), false alarms and misses affect automation dependence via two independent processes that manifest in categorically different behaviors: compliance and

reliance. Operator response to a warning signal is referred to as *compliance*. Operator non-action in response to a silent system indicating normal operation (i.e., in absence of a warning signal) is referred to as *reliance* (Chancey, Bliss, Yamani, & Handley, 2017). Meyer (2001, 2004) proposed that reliance and compliance are independent functions of false alarm rate and miss rate, respectively. While miss rate does seem to only influence reliance, excessive false alarms seem to degrade both reliance and compliance. Dickson and Wickens (2006) investigated the independence of compliance and reliance in a multitask environment, a simulated UAV task, and found that automation false alarms negatively affected both operator compliance and reliance. Wickens, Dixon, Goh, and Hammer (2005) found similar evidence using eye tracking metrics in the same UAV supervisory control task. Dixon, Wickens, and McCarley (2007) found that miss-prone automation did not affect operator compliance, but negatively affected concurrent task performance because the operator had to shift attention away from the concurrent task in order to catch potential automation misses. They also found that FA-prone automation negatively affected performance on the automated task due to reduced operator compliance, the “cry wolf” effect, and negatively affected performance on the concurrent task due to reduced operator reliance. The latter finding indicates that participants diverted their attention from the concurrent task to monitor the raw data in the automated task.

In general, in dual-task paradigms, FA-prone automation has been found to affect both operator compliance and reliance and, as a result, concurrent task performance as much, if not more, than miss-prone automation. FA-prone automation, on the other hand, negatively impacts performance on the automated task more than miss-prone automation due to the “cry wolf” effect (Dixon, Wickens, & McCarley, 2007; Levinthal & Wickens, 2006; Wickens, Dixon, & Johnson, 2005). However, the literature is inconsistent with regard to the ultimate effect of automation

bias on operator multi-task performance. As Levinthal and Wickens (2006) noted, many studies (e.g., Wickens, Dixon, Goh, & Hammer, 2005; Wickens, Dixon, & Johnson, 2005) involved physically demanding manual control of UAVs and, thus, a larger potential performance hit associated with task switching. In their study, which employed waypoint navigation and did not require participants to manually control the UAVs, miss-prone automation on a concurrent target identification task increased compliance and decreased reliance, and false-alarm prone automation decreased compliance and increased reliance. In addition, although Levinthal and Wickens (2006) found that false-alarm prone automation was potentially more disruptive due to its association with delayed response times to automation alerts consistent with the “cry wolf” effect, they ultimately found no effect of automation bias on concurrent UAV routing task performance. This is perhaps because the UAV task was less demanding than the full manual control required by Wickens, Dixon, Goh, and Hammer (2005) and Wickens, Dixon, and Johnson (2005), so participants could strategically allocate their attention between the automated target identification task and concurrent UAV routing task.

Automation dependence and the reliance-compliance dichotomization at higher LOAs. However, although the literature agrees that excessive false alarms affect user compliance, usually operationalized as the response time to automated alerts, there is a possibility that operators’ compliance behavior with higher LOA aids, such as SCOUT’s level six veto automation, might be qualitatively different than the frequently studied levels four and five since LOAs six and above do not necessarily require human interaction with the system (Table 2.2). Since human interaction with the automation is not required, response time is a poor indicator of automation compliance and reliance. In the instance of SCOUT’s automation, and other higher LOA aids, eye tracking measures could be used to gauge operator compliance and

reliance in absence of traditional performance metrics. Wickens, Dixon, Goh, and Hammer (2005) used visual scanning measures to assess user compliance and reliance on diagnostic automation during a simulated UAV supervisory control task. They found that miss-prone (60% reliable) automation reduced the percent dwell time (PDT) on the area of interest (AOI) representing the concurrent task and false-alarm prone automation (60% reliable) delayed the alert-driven shift in operator attention to the automated task AOI. The PDT that the eyes spent outside of the AOI representing the automated task was used as an indicator of reliance during periods of low workload. Visual scan response time (the time it took a participant's gaze to return to the automated task in response to an auditory alert) was used as an indicator of compliance during periods of high workload. However, this study utilized automation with a LOA of four (Table 2.2) and thus always required user interaction with the automated system.

In contrast, one would expect fundamentally different visual scanning behavior with level six and above automation, such as SCOUT's veto automation. While the PDT spent outside of the automated task AOI would remain a valid measure of reliance, visual scanning behavior indicative of compliance would look quite different. An operator complying with the automaton would not shift their attention to the automated task AOI in response to an auditory alert because user interaction is not required. Rather, a shift in visual attention to the automated task AOI would be more indicative of low compliance and low operator dependence on the automation. In the case of veto automation and higher LOAs, the PDT spent outside of the automated task AOI would also indicate operator compliance in addition to reliance. Therefore, in the case of veto automation and higher LOAs, the cognitive distinction between reliance and compliance becomes inconsequential because the resulting operator behavior looks the same when eye tracking metrics are used. In the case of level six veto automation, traditional performance

metrics could only be used to gauge what could be termed non-compliance, operationalized as the percentage of time an automatically selected target is overridden by the operator after an auditory alert (and perhaps the RT to override, though there are potential confounds associated with RT in a multi-task environment). However, this definition of ‘non-compliance’ is muddled with reliance because, fundamentally, the operator is engaging in such behavior because they hesitate to rely on the automation to indicate a problem on the automated task.

Therefore, when assessing operator interaction with veto automation and other higher LOAs, a more general measure of automation dependence is appropriate. Within SCOUT, automation dependence is operationalized as the percentage of responses that followed the automation’s recommendation. This is consistent with Calhoun et al.’s (2016) definition of automation dependence (which they termed “reliance”) for their study, in which they compared the effects of level five and level six automation on operator workload and performance on a UAV supervisory control task. Other studies involving veto automation exist (e.g., Liu, Wasson, & Vincenzi, 2009; Ruff, Narayanan, & Draper, 2002), but they are relatively uncommon. There appears to be a gap in the literature regarding the use of eye tracking data to examine operator dependence on automation at higher LOAs which, although not possible to employ in the present study due to technical constraints, is a potential topic for future research. Calhoun et al. (2016) attempted to use eye tracking metrics (including PDT on AOIs) to characterize operator dependence on level five and six automation, but experienced data quality problems.

Table 2.4

Summary of Low and High Reliability Operationalizations for Selected Automation Reliability Studies

Study	Low reliability	High reliability	Population	Task	Notes
Calhoun et al. (2016)	60%	86.7%	131 college students	UAV supervisory control	ALOA test bed used

Chancey, Bliss, Yamani, and Handley (2016)	60%	90%	88 undergraduate students	PC-based flight simulation with manual tracking and fuel management	Multi-Attribute Task Battery (MATB II) used
Dickson and Wickens (2006)	67% (Ex. 1) / 60% (Ex. 2)	100% (Ex. 1) / 80% (Ex. 2)	32 undergraduate and graduate students (including 20 licensed pilots)	UAV supervisory control	
Dixon, Wickens, and McCarley (2007)	60%	100%	32 undergraduate students	Concurrent two-dimensional continuous compensatory tracking task and system (single gauge) monitoring task	
Levinthal & Wickens (2006)	60%	90%	42 students	Concurrent UAV management and tank-detection task	SIL (Systems Integration Lab) UAV simulator used
Parasuraman, Molloy, and Singh (1993)	57.25%	87.5%	24 volunteers (10 men, 14 women; age: 19–43; right handed; corrected-to-normal vision)	PC-based flight simulation with manual tracking and fuel management and automated system-monitoring task	Multi-Attribute Task Battery (MAT) used
Rice (2009)	55%–95% in 5% increments		380 undergraduate students	Visual search task using still aerial photographs	Single-task environment; participant head position controlled with chin rest
Rovira, McGarry, and Parasuraman (2007)	60%	80%	18 undergraduate students	Low-fidelity command and control (C ²) sensor-to-shooter targeting simulation	
Ruff, Narayanan, and Draper (2002)	95%	100%	12 volunteers (8 men, 4 women; age: 22–49; corrected-to-normal vision; including two licensed pilots)	UAV supervisory control	UMAST (UAV Modeling and Analysis Simulator Testbed) ROV simulation software with UDAT (UMAST Decision Aiding Tool) used

Wickens, Dixon, Goh, and Hammer (2005)	60%	100%	39 student pilots	UAV supervisory control	
--	-----	------	-------------------	-------------------------	--

A paucity of small reliability manipulations. Furthermore, the majority of studies used unrealistically strong reliability manipulations, with two exceptions (Table 2.4). The first, Ruff, Narayanan, and Draper (2002), had only a 5% difference between their low and high automation reliability conditions. However, they found no main effect of reliability on operator performance, possibly due to limited power because of the relatively weak manipulation and/or the small sample size ($n = 12$). The second, Rice (2009), also employed a fine-grain reliability manipulation in a study wherein participants were tasked to locate a tank (or determine there was no tank). They were aided in their visual search by level four information analysis automation that varied in reliability from 95% to 55% in 5% increments. In addition to the reliability manipulation, the decision aid was biased either to produce only hits or only misses. However, while the results indicated a statistically significant increase in performance as the reliability of the decision aid improved, there was not a significant effect of automation bias on performance or a significant interaction effect for automation bias and reliability. While the primary purpose of Rice's study was to use a state-trace analysis to determine whether a multiple-process theory of operator trust could explain the effects of automation errors on automation dependence—and to that effect it did succeed in showing that FA-prone and miss-prone automation differently affected operator's reliance and compliance behavior—there was ultimately no effect of automation bias on operator performance. In addition, the study took place within a rigidly controlled single-task environment; visual angle and participant head position were controlled using a chin rest. While the high degree of control made sense for this particular study, there is still a current gap in the literature on the effects of smaller reliability manipulations on operator

performance and automation dependence-related behavior within a more realistic, complex, and noisy multi-task environment. It is still uncertain whether many commonly used (and some less commonly used, but promising) traditional and physiological performance metrics are sensitive enough to detect the effects of smaller reliability manipulations on operator performance and automation dependence-related behavior in a relatively noisy UAV supervisory control task.

The 4.5% difference in the low and high reliability conditions in the SCOUT study described herein is much smaller than the 30%+ difference in correct vs. incorrect automation responses commonly seen in the literature. This was an intentional experimental design decision; competing real-world decision aids will most likely not have such extreme differences in reliability, so it will be useful to see if the proposed metrics are sensitive enough to detect a weaker manipulation.

2.3 Trust in Automation

2.3.1 Trust from a stakeholder perspective. Lee and See (2004) define trust as “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” (p. 2). The proposed decentralized, flexible system of UAV control represents a significant change in how UAV operators must interact with autonomous systems, and what information they would require to support mission requirements. Underlying these considerations are two fundamental needs: the need to establish trust in automated UAV systems and the improvement of the trustworthiness of automated UAV capabilities. Trust in UAV automation must be built for initial stakeholders—system designers, testers, and policymakers—to feel confident fielding an UAV system; the competence of its automation must be high and its limitations known. Likewise, future automated UAV subsystems only have operational value if commanders and operators have sufficient confidence in their reliability and resilience to deploy

them on missions. Improvements in the trustworthiness of automated capabilities will thus increase operational value (DSB, 2016).

Unlike most commercial autonomous systems, which are designed for use in benign environments, autonomous systems designed for military application must be able to function in complex, unpredictable environments with the possible presence of adversaries intent on defeating their normal operation. In such high-stakes environments, it is critical that the operator be able to trust the automation. One barrier to trust is that automation lacks human-analog sensation, perception, and decision making. The different sensors and data sources that inform the automation's decision-making processes are not the same as those of its human operator, and it could therefore be operating on different contextual assumptions. Moreover, machine learning, reasoning, and decision-making can take vastly different paths to that of humans, which could lead human operators to question the trustworthiness of their machine partners (DSB, 2016).

The formation of human trust in automation begins at design time, with the establishment of what the automation can and cannot handle. Additionally, the system design should include real-time indicators of automation trustworthiness so that the operator can deal with variations in automation reliability when the operational environment exceeds the original design envelope or assumptions of the automated system. However, a basic awareness of systemic and/or environmental changes is not enough; the automation must be able to adapt to these changes. It must also effectively communicate changes in its own state and the environmental effects on its reliability to its human operator. System design should include sufficient anticipatory indicators so that the system is predictable and, should the environment exceed the design envelope of the automation, allows the operator to intervene in a timely and effective manner (DSB, 2016).

The building of stakeholder trust in new automation is yet another reason why a new suite of UAV operator performance metrics is needed: to establish the human performance impacts of new automation and its reliability, both to increase stakeholder buy-in and trust in the system and to inform the development of anticipatory aids and real-time indicators of automation reliability.

2.3.2 Trust an indicator of automation use: Do operators always notice poor reliability? Much of the literature postulates that automation reliability is an important factor of human use of automated systems because of its influence on operator trust; unreliable automation lowers human trust and can thus negate the benefits of implementing automation in the first place (Bliss, Gilson, & Deaton, 1995; Dixon & Wickens, 2006; Dixon, Wickens, & Chang, 2005).

As previously stated, Lee and See (2004) define trust as “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” (p. 2). While the “agent” can also refer to another person, for this discussion it shall refer to automation that interacts with the environment on behalf of the operator. Automated systems may be underutilized or disabled due to operator mistrust, as is the case with systems that give frequent false alarms (Parasuraman, Sheridan, & Wickens, 2000). There is evidence that false alarms are more damaging to operator trust than misses (Bliss, 2003) and that misses and false alarms affect operator trust differently (Dixon & Wickens, 2006; Dixon, Wickens, & McCarley, 2007; Meyer, 2001, 2004; Rice, 2009).

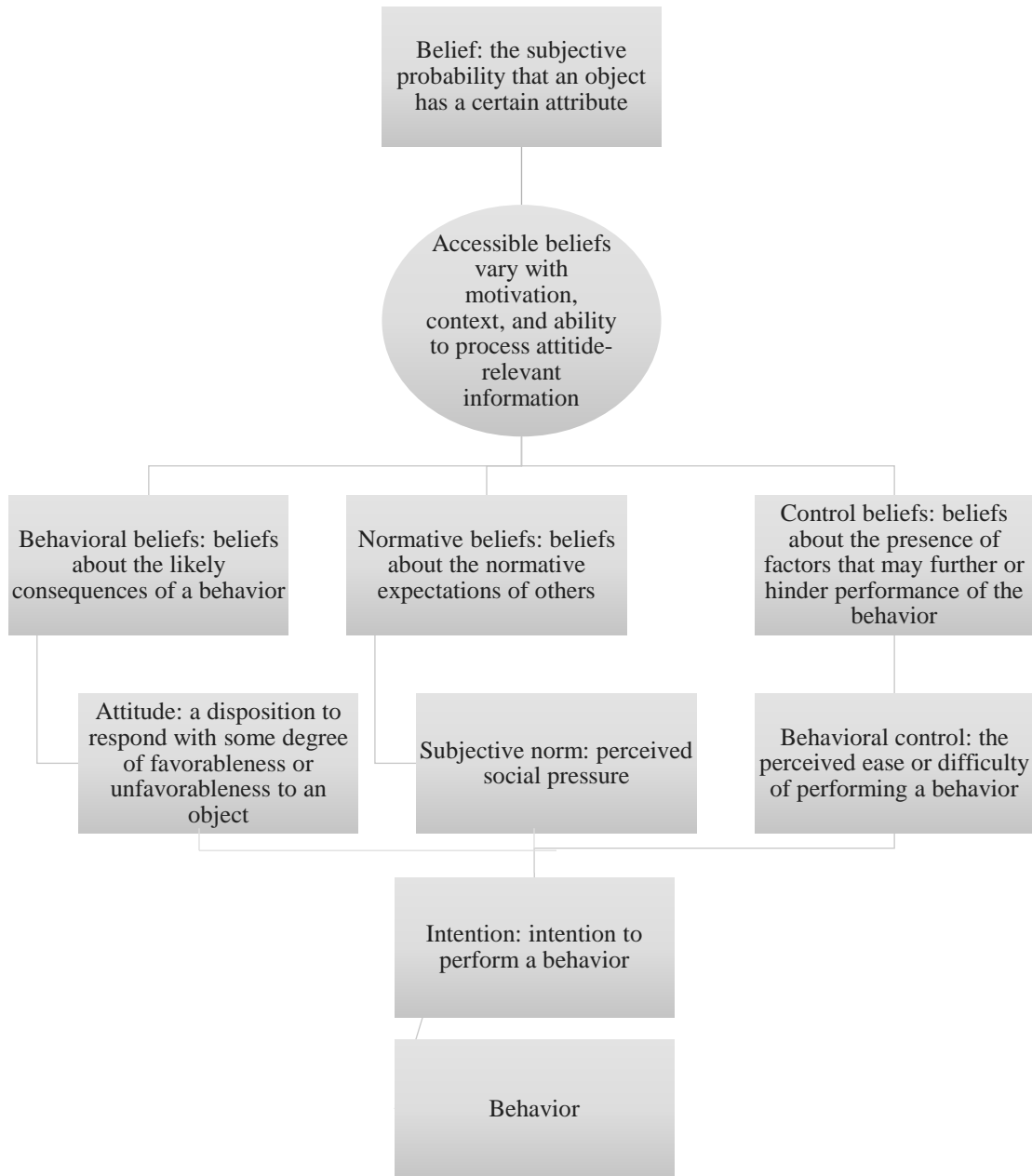


Figure 2.3. The relationship between beliefs, attitudes, intentions, and behaviors according to the theory of planned behavior. Adapted from “Attitudes and the Attitude-Behavior Relation: Reasoned and Automatic Processes” by I. Ajzen and M. Fishbein, 2000, *European Review of Social Psychology*, 11(1), 1–33. Copyright 2000 by Taylor & Francis.

However, these hypothesized independent types of trust, trust in signals and trust in non-signals, are inferred based upon two qualitatively different behavioral manifestations of automation dependence, i.e., reliance and compliance (Meyer, 2001, 2004), which are subject to

potential confounds. While some studies have shown that trust does affect automation dependence and can be measured consistently, it does not completely determine dependence. In fact, as Lee and See (2004) noted in their review of automation trust and dependence literature, the recent surge of studies has produced many confusing and seemingly conflicting findings. One issue is that the literature employs different operational definitions of trust and is inconsistent on whether trust is a belief, attitude, intention, or behavior. Lee and See suggest Ajzen and Fishbein's (1975, 1980) *theory of reasoned action* as a helpful framework to reconcile these conflicting definitions of trust. The expanded version of this framework, the *theory of planned behavior* (Ajzen, 1988, 1991), is presented in Figure 2.3. Within this framework, trust affects automation dependence as an attitude rather than a belief, and trust is not the sole mediating factor between operator beliefs about automation characteristics and their behavior. Defining trust as an intention or behavior invites the possibility of confounding its effect on surveyed intent or observable behavior with other factors, such as workload, situation awareness, perceived risk, and operator self-confidence (Lee & Moray, 1994; Lee & See, 2004; Parasuraman & Riley, 1997).

Lee and Moray (1992) found that, under certain conditions, operator reliance on automation did not correspond to changes in their trust. Their follow-up study provided evidence that, in addition to trust, operators' self-confidence in their ability to manually control an automated system predicted their dependence on the automation. They found that, in general, operators depend on automation when their trust in the automation exceeds their self-confidence that they could perform the task manually. Most operators in the study first adopted a predominantly manual control strategy and reported high self-confidence in their ability to manually control the system until a fault interrupted their manual control strategy. The fault led

to increased use of the automation, decreased self-confidence, and increased trust in the automated system. Furthermore, the operators tended to prefer either fully manual or fully automated control and their automation dependence displayed inertia (i.e., they were reluctant to change their automation use even when their trust and self-confidence changed). Individual biases also influenced the choice of fully manual or fully automated control, though operators generally preferred manual control.

However, the simulated process control plant “microworld” used in the study was much simpler than its corresponding commercial system. The authors noted that their findings might not scale to more complex systems, where a greater number of factors may influence operator reliance on the automation. For example, intermediate LOAs blur the distinction between fully manual and automatic control and thus may limit the generalizability of Lee and Moray’s (1992) study to environments that employ mid-level LOAs, such as SCOUT. In addition, the participants in Lee and Moray’s study were given a manageable task load so they could manually control all the pumps if they so desired. In a complex, time-pressured, multi-task environment, participants may not be able to complete all tasks without the assistance of an automated aid and may shed tasks to the automation due to time pressure.

Operator workload has been shown to affect operator reliance on automation. Because of this, while the traditional view of human factors professionals has been that humans should always have the ultimate decision-making authority in human-machine systems (e.g., Billings, 1991, 1997; Woods & Roth, 1988), Moray, Inagaki, and Itoh (2000) suggest that automation should have the final authority in time-critical situations, provided the performance payoff of using the automation outweighs the cost of potential automation errors, because an operator may not have time to engage the automation. The design of SCOUT’s level six automation is a

compromise between these two viewpoints. In absence of user interaction, the automation will carry out its recommended action, but the operator has access to the raw data and is able to override the automation within a limited time frame. This design is appropriate for the occasionally time-critical, relatively low-risk ISR mission environment.

Initial experimentation within the SCOUT environment has indicated that participants may not even be cognizant of the reliability of the automated aid and, thus, whether or not they can trust it (C. Sibley, personal communication, December 20, 2017). A survey based on Lee and Moray's (2004) validated subjective assessment of trust and self-confidence was administered as part of the present study to test this assumption. Operator task loads, and their ratings on the associated subjective workload metric, seem to be a far better predictor of automation dependence than their subjective ratings of trust in the automated aid. It is thus possible that, given a sufficiently complex and time-pressured multi-task environment, operators will depend on automation irrespective of its reliability and their trust in it. If this is the case, behavioral indicators of automation dependence, such as operator percent agreement with the automation, should increase during periods of high task load. In future studies, eye tracking metrics could also indicate increased automation dependence during periods of high task load; operators will spend less time (PDT) monitoring the AOI associated with the automated task even if their increased dependence is not reflected in their subjective trust and self-confidence ratings.

This hypothesis is consistent with the automation complacency literature. Automation complacency occurs in multi-task environments, when manual tasks compete with the automated task for the operator's attention. There is presently no consensus on the definition of complacency, but Parasuraman and Manzey (2010) cite the following core set of features between the various operationalizations: (1) human operator monitoring of an automated system

is involved; (2) the frequency of such monitoring is lower than some standard or optimal value; and (3) as a result of substandard monitoring, there is some directly observable effect on system performance, usually that a system malfunction, anomalous condition, or outright failure is missed (p. 382). Automation-related complacency has been implicated as a major contributing factor to aviation accidents (Funk et al., 1999).

A number of studies have used eye tracking metrics to examine attention allocation in systems under manual and automated control and found a relationship between automation complacency and reduced visual attention to the primary information sources feeding the automation that must be monitored (Baghieri & Jamieson, 2004; Metzger & Parasuraman, 2005; Wickens, Dixon, Goh, & Hammer, 2005).

However, evidence of automation complacency has also been obtained using more traditional performance metrics such as participants' response time and/or accuracy following an automation failure. These measures are easier to implement and are not subject to the data quality problems that often plague eye-tracking studies. Parasuraman, Molloy, and Singh (1993) operationalized complacency as the "failure to respond to an automation malfunction" (p. 17). More specifically, they characterized it as a combination of the mean probability of malfunction detection, the mean malfunction detection response time, and the number of false alarms made on the Multiple Task Battery (MATB) (Comstock & Arnegard, 1992). The MATB is multitask testing environment that includes a two-dimensional compensatory tracing task, an engine fuel management task, and an engine-monitoring task. The engine-monitoring task, which involves monitoring a cluster of four gauges to detect malfunctions, is supported by automation of variable reliability. They found that automation with consistent (as opposed to variable) reliability over time is more likely to induce complacency in a multitask environment. This

finding is consistent with Langer's (1989) concept of *premature cognitive commitment*, which she defined as the condition whereby one accepts and commits to "an impression or a piece of information at face value, with no reason to think critically about it" (p. 22). This rigid commitment that a person forms in response to conditions surrounding initial exposure to information can limit their subsequent use of said information (Chanowitz & Langer, 1981). In other words, participants exposed to automation of consistent reliability are more likely to develop a premature cognitive commitment about the efficacy of said automation and are thus more likely to become complacent. Participants exposed to automation of inconsistent reliability are less likely to develop a premature cognitive commitment and should possess a more open attitude toward the efficacy of the automation.

Parasuraman et al. (1993) found that the *consistency* of automation reliability over time was a greater contributing factor to participant complacency than initial reliability or absolute reliability level. On the other hand, it is possible that a statistically significant effect for absolute reliability level was not observed simply due to low power. Participants did, in fact, detect more automation failures in the low reliability condition, though the difference was not statistically significant. Bagheri and Jamieson (2004) replicated the study and found that participants detected significantly more automation failures when automation reliability was low.

Parasuraman et al. (1993) also found that high task load exacerbated automation complacency and detection of automation failures was significantly worse in a multitask environment. Their findings suggest that complacency is not a passive state into which an operator falls, but rather an active reallocation of attention away from the automated task to other manual tasks in instances of high workload (Parasuraman & Manzey, 2010). Complacency is reduced when the reliability of the automated system is low and variable, but still persists even at

low levels of reliability relative to manual performance (Bagheri & Jamieson, 2004; May, Molloy, & Parasuraman, 1993 as cited in Parasuraman & Manzey, 2010; Parasuraman, Molloy, & Singh, 1993).

2.4 Environment Complexity

Unmanned systems, like other complex systems, consist of many interacting components, the aggregate activity of which is nonlinear (Joslyn & Rocha, 2000). Nonlinear systems cannot be understood by studying each of their multiple subunits individually and treating the global system as the net value of the subunits. *Superposition* does not hold for nonlinear systems because the components of these systems interact (Goldberger, 2006). In other words, the whole of the system is not equal to the sum of its parts. “Agents residing on one scale start producing behavior that lies one scale above them. “[Just as] ants create colonies, urbanites create neighborhoods, [and] simple pattern-recognition software learns how to recommend new books,” numerous socio-technical components— stakeholders (e.g., operators and customers); vehicles and their subsystems; interacting friendly, hostile, and non-combatant entities; terrain features and manmade boundaries; weather and other atmospheric factors; operational environment characteristics; etc.—coalesce into the complex working environment of a UAV supervisory control operator.

This movement from low-level causality to higher-level sophistication is known as *emergence* (Johnson, 2012, p. 18). At each new level of complexity, new properties of the UAV supervisory control environment will emerge and additional research will be required to understand these unexpected characteristics and behaviors. In the words of Karl Marx, as a system increases in scale and complexity, “quantitative differences become qualitative ones” (as cited in Anderson, 1972, p. 396).

Phenomena can be described as *weakly emergent* or *strongly emergent* with respect to low-level causality. Weakly emergent phenomena are unexpected based upon the rules governing the behavior of low-level agents, but are nevertheless deducible in principle. In contrast, strongly emergent phenomena cannot be explained or predicted based upon the rules governing the behavior of low-level agents (Chalmers, 2006). The existence of strong emergence is controversial. Some believe that strong emergence is merely an artifact of humans' limited capacity to calculate and predict complex phenomena. However, regardless of whether strong emergence is existent or artifact, the point is that, at this time, it cannot be accounted for using reductionist methods (Pariès, 2006). It is therefore extremely unlikely that the aggregate activity of the future UAV supervisory control system will be fully predicable with respect to the rules governing its numerous socio-technical components.

Unfortunately, while the SCOUT environment as a whole is complex and subject to weakly emergent system behavior, one weakness of the current iteration is the lack of a manipulable subtask that produces emergent phenomenon in isolation. In future versions of SCOUT, weather might be incorporated into the route-planning task so that experimenters have a cleaner way to manipulate task complexity.

At present, however, the closest approximation to manipulating subtask complexity in SCOUT is the manipulation of the degree of uncertainty inherent in the payload (target identification) task. While the code behind the payload feed follows logical, predicable rules (i.e., the participant will not contend with emergent system behavior), the experimenter can still manipulate the amount of information, or number of target factors, that a participant must consider to effectively discriminate targets.

Although there is a substantial amount of literature on the relationship between automation reliability and operator performance, less attention has been paid to the role of task complexity, or even task difficulty. There is some evidence that task complexity or difficulty increases operator behaviors associated with automation dependence and decreases performance; the evidence is mixed, however, possibly due to the various unquantifiable ways that task complexity and/or difficulty is operationalized.

McFadden, Giesbrecht, and Gula (1998) found that operators were more reliant on an automated cue when the automated task became more difficult. Liu and Wickens (1987), on the other hand, did not find a significant effect of task difficulty on decision accuracy in either a single-task or dual-task environment. The study involved the manipulation of the difficulty of both a spatial and verbal decision task in single and dual task environments. The dual task environment incorporated a one-dimensional compensatory tracking task. The spatial task involved predicting future enemy position based on vectors that indicated current enemy position. In the difficult version of the task, both the direction and position of the vectors were relevant. In the easy version of the task, only the position of the vector was relevant. A concurrent verbal-arithmetic task also varied in difficulty. Only positive numbers were used in the easy condition, but the difficult condition utilized both positive and negative numbers.

Maltz and Shinar (2003), however, found that operator reliance on an automated aid correlated significantly with task difficulty. They also found that the automated aid interfered with performance on an easy task but aided performance on a more difficult task. In this study, participants were asked to locate military vehicles on still images of various terrains. The images were slightly blurred to increase uncertainty. The easy version of the task involved locating the vehicles in visible color images. The difficult version of the task involved locating the vehicles

on monochrome infrared images. Performance on the task was operationalized as an operator's probability of target detection and their false alarm rate. The performance of a control group was used to infer that locating vehicles within the monochrome infrared images was indeed more difficult than in the color image condition. However, the exact degree to which the tasks differed in difficulty is unknown.

In an effort to explicitly quantify the complexity of the SCOUT payload task, the complexity of target discrimination in the low and high complexity conditions will be calculated using Shannon entropy. Shannon's Information Entropy is a measure of information content in a given random sequence of information produced by a stochastic source of data. It could be thought as the uncertainty or unpredictability of the information in the data sequence. More specifically, for a discrete random variable X with possible values $\{x_1, x_2 \dots x_n\}$ and probability mass function $P(x_i), i = 1, 2, \dots, n$, the Shannon entropy is defined as in Equation 1.1 below (Teixeira, Matos, Souto, & Antunes, 2011).

$$H(X) = - \sum P(x_i) \log_2 P(x_i) \quad (1.1)$$

Information entropy can also be used as a measure for the complexity contained in a sequence data of a random variable. Although different from Kolmogorov Complexity in terms of calculation, there are some equivalencies between these two measures; the expected value of Kolmogorov complexity equals Shannon entropy, up to a boundary (Teixeira, Matos, Souto, & Antunes, 2011).

2.5 Task Load and Subjective Workload

In the experiment described herein, participants are exposed to variable task load and its effects on performance are assessed. For the purposes of this study, task load is defined as the number of tasks assigned to an operator per unit of time. This definition is conceptually similar

to utilization, or the percent of time an operator is busy. Cummings and Nehme (2009), who defined workload as utilization, found evidence of a parabolic utilization-performance curve analogous to the Yerkes-Dodson relationship. While the original Yerkes-Dodson curve and its associated research focused on stimulus strength and learning, a similar relationship between arousal and performance was later identified (Hebb, 1955).

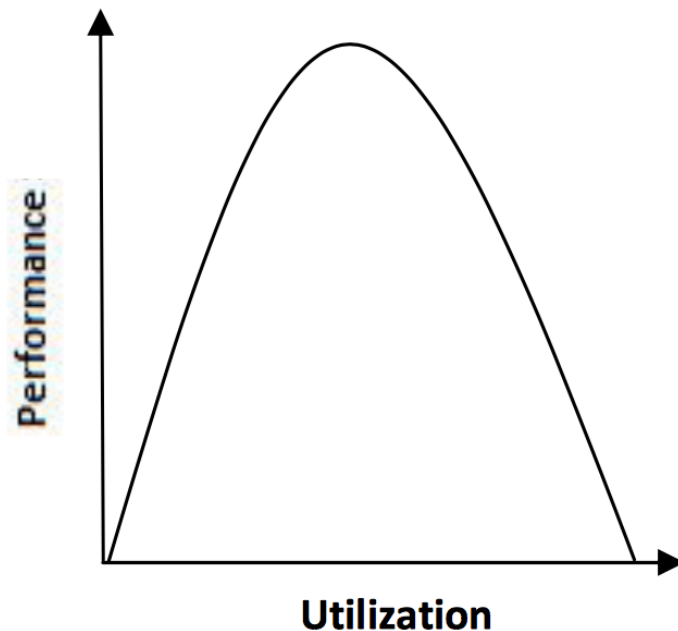


Figure 2.4. Workload-performance relationship. Reprinted from “Modeling the Impact of Workload in Network Centric Supervisory Control Settings,” by M. L. Cummings and C. E. Nehme, 2009, 2nd Annual Sustaining Performance Under Stress Symposium.

In general, Cumming and Nehme’s (2009) utilization-performance curve indicates that operators perform best under moderate levels of utilization. High and low levels of utilization will degrade performance (Figure 2.4).

While task load (defined as the number of tasks per unit time) and utilization are relatively cut-and-dry concepts, the factors contributing to operators’ subjective experiences of mental workload are not as straightforward. In fact, there is no universally accepted definition of workload despite decades of research on the subject (Cain, 2007). Proposed operational

definitions continue to disagree about its source(s), mechanism(s), consequence(s), and measurement (Huey & Wickens, 1993). The proposed aspects of workload seem to fall within three broad categories: the amount of work and number of things to do, time and the particular aspect of time one is concerned with, and the subjective psychological experiences of the human operator (Cain, 2007; Lysaght, Hill et al., 1989).

One common definition of subjective mental workload is a "participant's direct estimate or comparative judgment of the mental or cognitive workload experienced at a given moment" (Luximon & Goonetilleke, 2001, p. 230). Subjective mental workload can be assessed using a variety of rating techniques. Arguably the most ubiquitous of these ratings is the NASA Task Load Index (NASA-TLX). The NASA-TLX assesses workload across six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration (Hart & Staveland, 1988). Twenty-step bipolar scales are used to obtain a score, which ranges from 0 to 100, for each dimension. Then, in a pairwise comparison task, the participant selects which of the six dimensions are most relevant to workload in the task being measured across all pairs of the six dimensions. The results of this task then determine the weighting each of the six dimensions receives when they are combined into a global score (Rubio et al., 2004; Schnell et al., 2014). While the NASA-TLX has good reliability and validity, the pairwise comparisons increase the time and effort needed to administer the rating technique.

Due to time constraints and the desire for an absolute assessment of workload rather than a relative assessment of different workload factors, the experimenters chose not to use the NASA-TLX to assess subjective workload. Instead, the Crew Status Survey (CSS) was employed.

The CSS was originally developed by the Air Force School of Aerospace Medicine (Samn & Perelli, 1982) and later revised and verified by the Air Force Flight Test Center (Ames & George, 1993). It is designed to reduce time for crews in a field research setting to report subjective workload and fatigue data. The CSS consists of two seven-point, forced-choice Likert-type scales. Each point on both scales is anchored and higher numbers indicate a greater feeling of subjective workload or fatigue (Samn & Perelli, 1982). The CSS assesses four components of subjective workload: activity level, system demands, time loads, and safety concerns. However, although the scale anchors reflect a multidimensional concept of workload, raters must implicitly integrate the various workload factors into a unidimensional workload rating ranging from one (least workload) to seven (most workload) (Ames & George, 1993).

According to Ames & George (1993), the CSS is appropriate for use in situations where an absolute assessment of workload is needed (rather than a relative assessment), where an easy to understand scale is needed, where minimal participant training time is available, and where the collected data may be analyzed using statistical procedures requiring interval quality data.

It is important to note, however, that performance and subjective measures of workload can dissociate under certain conditions, such as when an operator invests greater resources to improve their performance of a resource-limited task (i.e., they try harder). In multitask environments, such as SCOUT, time-sharing between concurrent tasks or between display elements could also place additional demands on working memory (Yeh & Wickens, 1988). Therefore, the effects of task load on performance might not be mirrored by its effect on subjective workload ratings or, if utilized, common physiological indicators of mental effort (e.g., heart rate variability or pupil diameter). If an operator tries harder than average on the

SCOUT task and performs well, their subjective workload ratings might be high, but so will their level of performance.

2.6 Situation Awareness and UAV Ground Control Station Design

The future UAV supervisory control paradigm envisioned by the DoD will require a single operator to control multiple UAVs. These UAVs will be semi-autonomous, meaning that they will have the capacity to make certain higher-order decisions independent of operator input and predefined mission plans. This means that, while the “stick and rudder” tasking of a UAV operator might decrease, their new supervisory role introduces a new source of workload in the form of rapid judgment of the appropriateness of decisions and actions made by the automation and the projection of their impact on overall mission objectives. Operators, who must monitor an increasing number of automated systems, will thus be challenged to remain “in the loop” through long periods of nominal operations while remaining poised to engage in short bursts of time-sensitive contingency operations when the automated systems encounter a situation beyond the scope of their design and either respond inappropriately or fail (Ruff et al., 2004). In other words, human supervisors of highly automated systems often struggle to intervene in system control loops and assume manual control when environmental conditions exceed the design of the automation; the resulting situation awareness and performance decrement is known as out-of-the-loop (OOTL) performance (Kaber & Endsley, 2003; Kessel & Wickens, 1982; Young, 1969).

Unfortunately, there are documented instances of increases in the automation of manned systems causing significant fluctuations in operator workload, loss of situation awareness, and decrements in performance (i.e., in OOTL performance). According to Parasuraman, Sheridan, and Wickens (2000), issues associated with automation management include task allocation

between operator and system, human vigilance decrements, clumsy automation, limited system flexibility, mode awareness, trust/acceptance, failure detection, and automation biases.

UAV ground control stations, like most modern operator interfaces, can produce large amounts of data on both the status of various subsystems and the external environment. The problem with modern operator interfaces is rarely a paucity of information. Rather, poorly designed UAV ground control stations (GCSs) and other supervisory control interfaces are notorious for inundating operators with data while making it difficult to find the information they need for good task performance and decision-making (Endsley, 2000). For example, the human factors problems plaguing the Predator GCS were succinctly described by Col. John Dougherty, an MQ-1 Predator operations commander with the North Dakota National Guard: “Too many screens with too much information, folks.” The predator GCS, which was originally a technology demonstration project, was rushed into widespread use once its value became apparent. However, because of its rushed development cycle and requirements creep, subsystems were added piecemeal, each with its own unique, user-unfriendly display window (Freedberg, 2012). The subsystem windows can be opened on top of each other, resulting in substantial display clutter and reduced salience of mission-critical information. It is thus difficult for operators to locate information needed for decision-making under dynamic operational constraints, let alone correctly integrating and interpreting said information. It is becoming widely recognized that “more data does not equal more information” and the indiscriminate introduction of automation and “intelligent systems” generally exacerbates degraded operator SA and OOTL performance rather than mitigating it (Endsley, 2000; Endsley & Kiris, 1995).

The enhancement of operator SA is a major design goal for developers of operator interfaces, automation concepts, and training programs (Endsley, 2000). Ruff et al. (2004)

propose using multiple levels-of-automation (LOAs) to keep the operator “in the loop” for optimal SA, workload, and decision-making during a supervisory control task. The pervasiveness of automation can vary across a continuum of levels ranging from no automation (i.e., fully manual performance by the human operator) to completely automated systems that require no human input during nominal operations. While higher LOAs might allow a single operator to control more vehicles, they also tend to remove the operator from the loop and can result in poor performance in response to automation errors. While more intermediate LOAs would limit the number of UAVs a single operator could control, Ruff et al. (2004) hypothesize that such “an intermediate LOA could improve performance and SA, even as system complexity increases and automation fails” (p. 219). Some research supports this hypothesis (e.g., Ruff, Narayanan, & Draper, 2002), while others (e.g., Endsley & Kaber, 1999) cite additional factors that can impact the benefit of LOA, such as whether the task involves option selection versus higher-level cognition. These results indicate a need for more research investigating LOAs in different task environments (Ruff et al, 2004).

Situation awareness (SA) is a ubiquitous concept that is often discussed in both the commercial and military aviation communities and the human factors field as though its meaning were self-evident. However, as a psychological construct, SA is not readily observable and is thus difficult to operationally define. As one may expect, there is no definitive operational definition of SA and many divergent—and sometimes even conflicting—definitions of SA have surfaced in the literature (Adams, Tenney, & Pew, 1995; Uhlarik & Comerford, 2002). Some of these definitions are more general, while others are more domain-specific (Dominguez, 1994; Endsley, 2000).

Just as there are many ways to operationally define SA, there are many ways to measure it. These measures, which are based upon different theoretical constructs of SA and thus utilize different operational definitions of SA, can be divided into three general categories: (a) explicit (e.g., SAGAT); (b) subjective (e.g., SART, SA-SWORD); and (c) implicit measures of SA (e.g., PPI) (Schnell et al., 2014; Uhlarik & Comerford, 2002).

Endsley's (1988a) definition of SA is a well-accepted general definition that has been found to be applicable across a wide variety of domains. According to this definition, SA is "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" (p. 97).

The first level of SA, "perception of the elements in the environment within a volume of time and space," involves the perception of the status, attributes, and dynamics of relevant elements in the environment. A UAV operator, for instance, might perceive the status of their assigned vehicles, targets of interest, environmental features such as elevated terrain or restricted airspace, and any relevant characteristics of these features (e.g., allegiance, location, capabilities, speed, shape, size, and color). The second level of SA, comprehension of the current situation, involves an operator integrating and interpreting the disparate elements perceived in level one to form a holistic picture of the environment and an understanding of significance of objects and events in regard to mission goals. For example, a UAV operator on an ISR mission might note a cluster of high-priority targets of interest and recognize the implications for that geographical location and its importance to enemy objectives. The third level of SA, projection of future status, refers to an operator's ability to project the future actions of the elements in the environment in the near term. For example, a UAV operator might recognize a pattern in the characteristics of the targets of interest that they are imaging and their location relative to

geographical features and other areas of interest. They can then use this information to anticipate the location of future targets so they can effectively allocate their assigned UAVs (which might have different speeds, capabilities, and ranges) to meet mission goals (Endsley, 1995).

2.7 The Need for New Operator State and Performance Metrics

The DoD and its NATO allies are working toward developing a Common Control Station (CCS) to replace existing stove piped, proprietary UAV GCSs, which are more costly (due to redundant procurement and training efforts) and limit innovation. The CCS, and other future ground control stations, will employ a service-oriented architecture (SOA), or a modular UAV control design that enables services to be easily replaced (Chanda et al., 2010; Sibley, Coyne, and Morrison, 2015). NATO's proposed functional architecture for UAV control systems and its required communication protocols are outlined in Standardization Agreement (STANAG) 4586 (NATO, 2012). STANAG 4586 also discusses the need for interface standardization, but stops short of specifying how the common control interface should look. The DoD released a style guide to provide system designer recommendations for how to display information within a UAV control station (OSD, 2012). However, it is still uncertain what information needs to be displayed, which is of critical concern as more systems become automated and humans are moved further OOTL (Sibley, Coyne, & Morrison, 2015).

The transition to multiple-UAV supervisory control will require a suite of new capabilities; these include better data visualization and decision support, alerting, and monitoring tools. These new automated tools, as with all proposed automated UAV subsystems, must be robust and their effects on the system predictable. Their actions must be clear and directly observable by human operators. They must possess sufficient self-awareness to know when they are operating at or near the limits of their design assumptions or operational boundaries, and they must be capable of providing real-time estimates of their reliability in response to dynamic

mission conditions. All these factors are critical to the establishment of operator trust in any new system, capability, or tool. A comprehensive set of metrics needs to be identified in order to adequately assess the potential benefits and costs of these new technologies. This set of metrics is especially important since novel capabilities are likely to be introduced over time.

Each year the DoD funds new tools to improve warfighter performance but, despite large investments, their operational utility is often questioned. Operator and mission performance metrics must be identified to quantify the impact of new tools on mission success and operator performance. Performance metrics within UAV operations are dependent on the mission context (e.g., phase of flight and mission priorities). Without the use of carefully operationalized and documented mission performance metrics and a common nomenclature for documenting the specific mission context, accurate comparison across different UAV team control structures or system interfaces would not be possible (Coyne, Sibley, & Morrow, 2015).

However, one caveat that should be noted regarding the current study and other UAV performance research efforts that employ traditional performance-based measures of accuracy and response time is that such measures will provide only a partial assessment when evaluating human performance issues in supervisory control tasks. This is because multiple-UAV control, like most supervisory control tasks, involves extended periods of monitoring where traditional performance data are not available. In other words, response time and accuracy measures are not available, let alone representative of good operator performance, when an operator spends the majority of the time their UAVs are enroute or loitering over areas of interest monitoring their moving map and payload feeds. Good performance under these conditions involves the maintenance of SA rather than direct interaction with the system. The development of performance measures for monitoring periods is an especially important topic for future research

considering the increased likelihood of degraded SA given the future unmanned vehicle control paradigm of increased automation. Studies have shown degraded SA can increase the time it takes an operator to re-engage with a system and react to sudden, critical mission events (Coyne, Sibley, & Morrow, 2015; Endsley & Kaber, 1999).

2.8 Supervisory Control Testing Environments

Two of the current challenges within supervisory control research for unmanned vehicles are that multiple UAV supervisory control systems do not yet exist within the DoD and the future Concept of Operations (CONOPS) is not well defined. Since concurrent control of multiple UAVs does not yet exist in any operational context, the research community has developed several test beds to simulate some of the different types of tasks an operator might have to conduct; the two most frequently used platforms are the Adaptive Levels of Automation (ALOA) and the Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles (RESCHU) test beds (Johnson, Leen, & Goldberg, 2007; Nehme, 2009). These tools have provided some valuable initial information on some of the potential benefits and challenges associated with different types and levels of automation within supervisory control (e.g., Calhoun, Draper & Ruff, 2009).

The Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles (RESCHU) is an online UAV and unmanned underwater vehicle (UUV) supervisory control test bed developed by the Human and Automation Laboratory at MIT. RESCHU's simulated ground control interface consists of a map display, camera window, vehicle control panel that displays vehicle health and mission information, and a mission timeline that gives the estimated time of arrival to areas of interest.

The RESCHU test bed is particularly valuable for research focused on how vehicle team heterogeneity affects operator performance. In RESCHU, operators can control a team consisting of up to three types of vehicles: a high altitude long endurance (HALE) UAV, a medium altitude long endurance (MALE) UAV, and a UUV. The vehicles have variable speeds (UUVs are slower than UAVs) and capabilities (HALE UAVs are used to locate new targets within an area of interest, while MALE UAVs and UUVs are used to acquire these pre-determined targets) (Nehme, 2009).

In addition, RESCHU can be used to conduct research focused on trust in automation since it employs a sub-optimal route planner. The route planner, by sometimes failing to assign the best paths and vehicle-target assignments, seeks to replicate the performance of real-world automation and serves as an additional source of operator workload since operators must reassign vehicles.

Moreover, two different versions of RESCHU were recently employed to assess the effect of UAV control architectures on operator workload and performance. The vehicle-based RESCHU interface employs a centralized control architecture in which a single operator individually tasks multiple UAVs. The task-based RESCHU interface employs a decentralized architecture that requires the operator to convey high-level goals (i.e., a task list) to an automated mission and payload manager, which then decides how to best to distribute the tasks among multiple UAVs. In general, decentralized control schemes are favored because they eliminate the UAV operator and their ground control station as a single point of system failure and are more robust to delayed operator action and lapses in situation awareness. However, decentralized control schemes are generally less resilient to unexpected events and emergent system behavior.

Given the limitations of both control architectures, it is likely a hybrid mix will be best for operational use (Cummings, Bertucelli, Macbeth, & Surana, 2014).

Table 2.5

Sheridan and Verplank's 10 Levels of Autonomy and their Availability in ALOA

Level	Description of System Output	Type	ALOA Task(s)
10	The computer decides everything, acts autonomously, ignoring the human	Fully Automatic	Weapon release authorization, image analysis, allocation, and autorouting
9	Informs the human only if it, the computer, decides to		
8	Informs the human only if asked		
7	Executes automatically, then necessarily informs the human	Automatic with feedback	Weapon release authorization, image analysis, and autorouting
6	Allows the human a restricted time to veto before automatic execution	Veto	Weapon release authorization, image analysis (single and multiple options), and autorouting (single and multiple options)
5	Executes that suggestion if the human approves	Consent	Weapon release authorization, image analysis (single and multiple options), and autorouting (single and multiple options)
4	Suggests one alternative		
3	Narrows the selection down to a few	Multiple options	
2	Offers a complete set of decision/action alternatives		Image analysis and autorouting
1	Offers no assistance; human must take all decisions and actions	Manual	Weapon release authorization, image analysis, allocation, and autorouting

Note. Adapted from (1) “A Model for Types and Levels of Human Interaction with Automation,” by R. Parasuraman, T. B. Sheridan, and C. D. Wickens, 2000, *IEEE Transactions on Systems, Man, and Cybernetics- Part A: Systems and Humans*, 30(3), p. 287. Copyright 2000 by IEEE. (2) “Testing adaptive levels of automation (ALOA) for UAV supervisory control (No. AFRL-HE-WP-TR-2007-0068),” by R. Johnson, M. Leen, and D. Goldberg, 2007. Copyright 2007 by the Air Force Research Laboratory.

The ALOA test bed was designed by the Air Force Research Laboratory to assess the effect of a range of levels of autonomy on an operator’s multiple-vehicle supervisory control performance. The levels of autonomy implemented in ALOA are based on Sheridan and Verplank’s 10 Levels of Autonomy (Table 2.5). Within the ALOA test bed, the level of automation for four

tasks can be set by the experimenter, dynamically controlled by the operator, or automatically adapted by the system in real time according to a workload-based, performance-based, or time-based technique; the four automated tasks include: weapon release authorization, image analysis, task allocation, and autorouting.

One strength of the ALOA test bed is that its design leverages UAV controller interview data to help operators maintain situation awareness of the mission, vehicle status, and environment. The ALOA interface includes a chat window that presents the rules of engagement (ROE) and mission updates, a scrolling ticker that displays warnings and system updates, color-coded vehicle Health and Status Indicators, a map display, and visual and aural Pop Up Threat Indicators. ALOA also includes planning tools to help users decide on a route; reallocate tasks; assess potential impacts of new threats; and avoid pop-up threats, such as surface-to-air missile (SAM) shots (Johnson, Leen, & Goldberg, 2007).

One of the limitations of the RESCHU and ALOA test beds, however, is that the tasking was developed to be quickly learned and tested on untrained populations. As such, the complexity is lacking in some of the tasking and is not especially representative of the tasks a current or future operator would be performing (i.e., decision making under uncertain contexts). Additionally, most supervisory control research has focused on scenarios with sustained high levels of workload where participants complete six to seven tasks per minute (e.g., Kidwell, Calhoun, Ruff & Parasuraman, 2012). For this reason, although the high degree of LOA control within ALOA makes the test bed very useful for investigating future adaptive automation strategies, its reliance on pop-up threats limits its ecological validity. As previously discussed, the task demands for current UAV operators are highly variable and increased automation leads to significant downtime during certain mission phases. Likewise, future multiple-vehicle supervisory control operators are also

expected to experience significant downtime due to increased automation. Similarly, the goal of RESCHU's surveillance-type missions is to detect and identify as many targets as possible. The mission performance metric used is the total number of correctly identified targets normalized by the total number of possible targets for the mission. This consistent level of tasking provides a near continuous measurement of performance that, while ideal for research, does not reflect the real environment. This task level only represents a narrow range of UAV mission contexts. There are many contexts in which a UAV operator will have limited interaction and must sustain their attention and SA for extended periods of time.

Assessing levels of automation and display formats within a single mission context limits the generalizability of the supervisory control research results to future operations. To apply existing scientific knowledge of supervisory control towards future systems, it is essential to assess tools and concepts within realistic, synthetic environments that can model the broad range of scenarios and contexts an operator would actually encounter (e.g., denied/degraded communications, sustained monitoring, and target-asset allocation under uncertain conditions).

2.9 The Development of the Supervisory Control Operations User Testbed (SCOUT)

The Naval Research Laboratory (NRL) developed SCOUT to begin to address some of these research needs. SCOUT was iteratively designed based on input and feedback from current UAV operators. UAV operators at Yuma proving grounds in Arizona were interviewed and asked to describe the challenges, common errors, and system abnormalities that they experienced while controlling contemporary UAVs. In addition, they were asked to envision future UAV supervisory control operations, with emphasis placed on how the aforementioned challenges, errors, and abnormalities might manifest in this environment. SCOUT was designed to abstract out the components of contemporary UAV control and to represent some of the challenges faced

by current UAV operators in order simulate the tasks in which a future UAV operator might engage while supervising multiple vehicles.

During a SCOUT scenario, participants manage three heterogeneous helicopter UAVs. In order to meet mission goals, they must decide how to best allocate the UAVs to locate targets while simultaneously completing a number of subtasks, including maintaining communication with command and intelligence personnel via chat, updating UAV parameters, and monitoring their sensor feeds and airspace. Points are assigned to various actions based on their mission priority and the goal is to obtain as many points as possible.

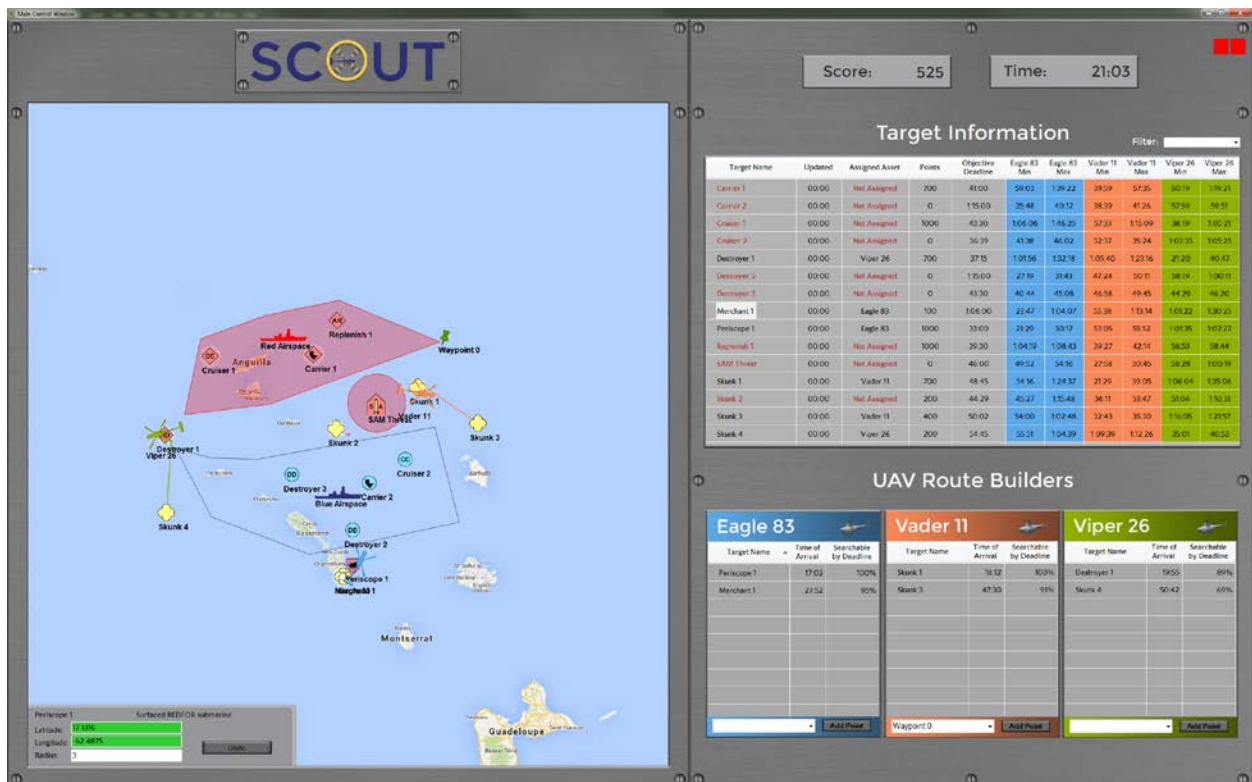


Figure 2.5. SCOUT route planning (left) screen.

SCOUT is available in both single-monitor and dual-monitor configurations. In the dual-monitor set-up, the left screen is primarily used for route planning (Figure 2.5). The Target Information table and UAV Route Builder boxes provide operators with estimated search times for each target, their point values (which indicate mission priority), their deadlines, the size of

their search areas, and the percent of those areas that can be covered by each UAV before the target deadlines. Each SCOUT mission involves a variable degree of uncertainty. Operators do not know the exact location of the targets within their search areas. They might find a target after searching only 1% of its search area, but they could also be required to search 100% of its search area to locate it. Moreover, the entire search area might not be traversable by the target deadline (when the intelligence expires and the location estimate becomes too uncertain to be useful). A SCOUT mission with a greater degree of uncertainty would generally involve targets with large search areas and short deadlines.

Additional sources of uncertainty include whether operators will be granted access to restricted operating zones (ROZs), which are indicated by the outlined and/or red-shaded areas on the moving map display, and the closeness of distractor targets to the actual target on the simulated payload task.

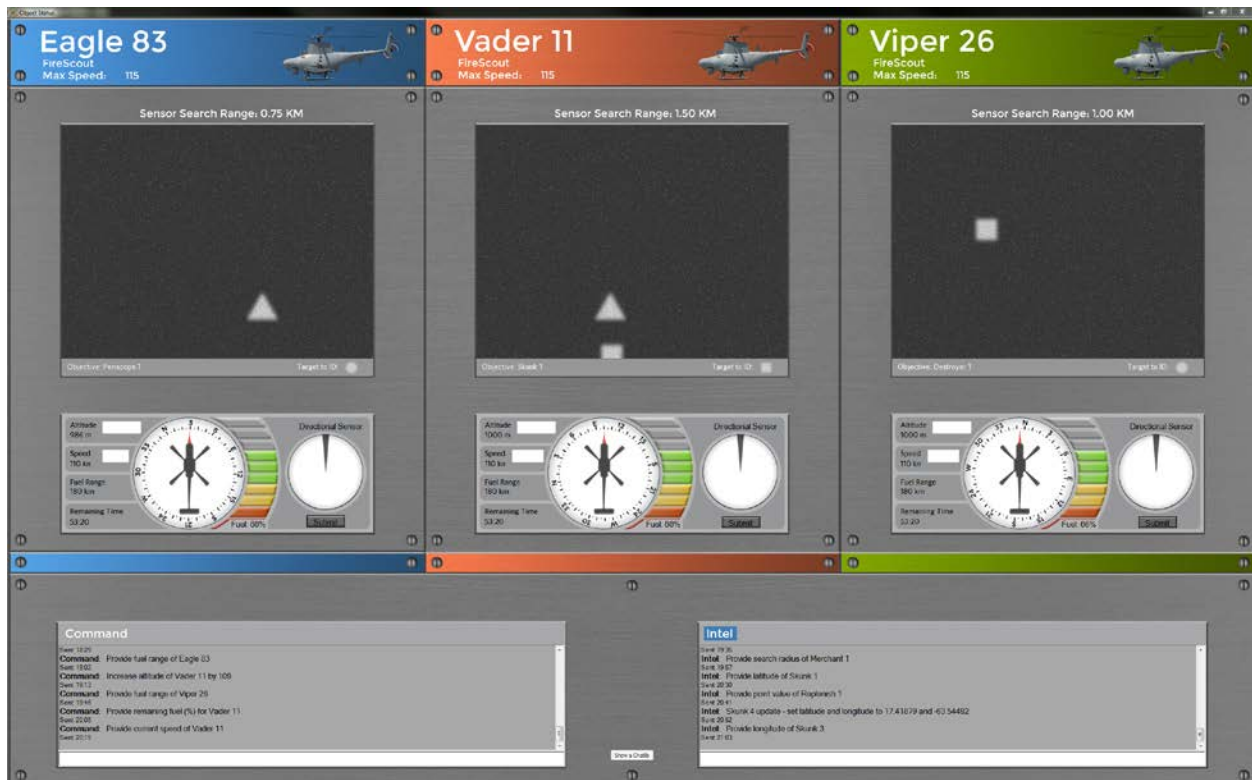


Figure 2.6. SCOUT vehicle status (right) screen.

The simulated payload task is located on the right screen along with other vehicle-centric information, such as fuel status, altitude, and speed. In the sample mission shown in Figure 2.6, Eagle 83 is searching for Periscope 1, which looks like a circle. In this case, the distractor targets (triangles and squares) are quite distinct from the target of interest. Additional uncertainty and complexity could be introduced into the scenario by using distractor targets closer in appearance to the actual target.

The complexity of a SCOUT scenario can be further altered by changing the degree of heterogeneity between the vehicles, increasing or decreasing the number of targets and/or variety of targets types on the moving map display, designing scenarios where there is or is not an obvious ideal route, and increasing or decreasing the overall detail of the payload task and the number of dimensions upon which targets and distractor targets differ. The high degree of flexibility in determining mission uncertainty and complexity, and the ability to create time pressure using target and message-response deadlines, make SCOUT an ideal test bed in which to study UAV operator performance, decision-making, and risk-taking under realistic operational conditions: complex, information-rich, and sometimes time-pressured.

In upcoming versions of SCOUT, a decision support tool will be available to help operators plan their routes. The support tool will consider the acceptability of risk for a given mission when deciding on a route plan to present for operator approval.

SCOUT can also be used to study operator behavior, SA, and performance in response to variable automation reliability and the resulting trust-in-automation issues that could arise. The payload task is equipped with level six (veto) automation with customizable hit and false alarm rates. When enabled, the automation highlights potential targets and, after giving the operator time to deselect erroneous selections, selects said targets. Since selecting an incorrect target (e.g.,

a circle instead of a square) results in lost points, reliance upon automation with a liberal response criterion could result in a significant point loss. On the other hand, reliance on automation with a conservative response criterion could result in the operator missing a target altogether.

Behind the scenes, SCOUT gathers and synchronizes all task/mission performance data with detailed information on the user's behavior and interactions with the system. If desired by the experimenter, SCOUT can also synchronize physiological data, such as eye gaze data, pupil size, heart rate, and respiration rate. SCOUT currently supports SmartEye Pro, GazePoint, EyeTribe, and Tobii EyeX eye tracking systems (Sibley, Coyne, & Thomas, 2016).

2.10 Literature Review Summary and Identified Gaps

In summary, there is a substantial body of literature on the effect of automation reliability on operator performance in single and multi-task environments. However, most of the existing research involves manipulating the reliability of level five information acquisition and analysis automation or lower (Parasuraman, Sheridan, & Wickens, 2000; Sheridan & Verplank, 1978). Few research efforts have focused on the effects of decision and action selection or action implementation automation of level six or higher on operator performance (e.g., Calhoun et al., 2016). Implementation of mid-level automation is likely, if not necessary, to enable supervisory control of multiple UAVs by a single operator in the relatively low-risk and predictable ISR mission environment.

Since level six (veto) automation does not require manual user confirmation of each selection, operational definitions of operator performance and automation dependence common to the literature must be reconsidered; certain prevalent performance metrics, such as response time, are not necessarily indicative of performance and/or automation dependence. The work

described herein seeks to build on the work of Calhoun et al. (2016) and further develop performance metrics that more accurately reflect operator performance in a supervisory control environment involving veto automation.

In addition, most existing studies involve artificially large reliability manipulations. While Rice (2009) and Ruff, Narayanan, and Draper (2002) implemented small reliability manipulations, their studies took place within a highly controlled single-task environment and were limited in power due to a small sample size, respectively. Moreover, since the purpose of an ISR mission is to image targets and the risk associated with a false alarm is relatively benign, system designers are much more likely to set the beta (i.e., alert threshold) so that the automation has a more liberal response criterion. Most existing studies characterize ‘unreliable’ automation as either unrealistically conservative (i.e., miss-prone) and/or generally unreliable to the point where it would likely not be used operationally (e.g., 60% reliable).

Research is needed to develop metrics that are sensitive enough to discriminate between the impacts of automated aids with smaller, more realistic reliability differences on human operator performance, as such metrics will be useful for the development, testing, and evaluation of future automated aids for UAV ground control stations.

Furthermore, while many studies have looked at the effect of trust on automation dependence, anecdotal evidence from prior studies using SCOUT indicate that its multitask environment is sufficiently complex and time-pressured enough that operators may depend on the payload task automation irrespective of its reliability and their trust in it. In fact, anecdotal evidence also indicates they might not even be cognizant of the automation’s reliability. This anecdotal evidence is consistent with the automation complacency literature and the study herein attempts to formally establish whether participants’ subjective trust and self-confidence ratings

reflect their degree of automation dependence. Based on initial anecdotal findings, and the findings of related studies, task load is likely a better predictor of automation dependence in multitask environments.

In contrast to the effects of automation reliability and task load on operator performance, trust, and automation dependence, relatively little attention has been paid to the effects of task environment complexity on operator performance (e.g., Maltz & Shinar, 2003). In addition, prior studies involved the comparison of qualitatively different low and high complexity tasks and did not seek to explicitly quantify the difference in complexity between experimental conditions. The research described herein seeks to expand on this limited body of research and investigates the use of Shannon entropy as a means to explicitly quantify the complexity of a search task.

This study also investigates the effect of task load on participants' subjective workload and fatigue ratings. Subjective workload could dissociate from traditional performance metrics if operators invest greater resources to improve their performance of a resource-limited task, in other words, if they try harder (Yeh & Wickens, 1988).

Perhaps most importantly, while the majority of previous studies used undergraduate students, the general population and, less commonly, civilian pilots, this research investigates the effects of task load, environment complexity, and automation reliability on UAV supervisory control performance within a very unique population: student naval aviators and naval flight officers.

In conclusion, the research described herein attempts to address the following identified gaps in the literature: (1) the effects of task load, environment complexity, and automation reliability on operator performance, which have not been investigated together; (2) the effects of veto automation reliability on operator performance and the unique challenges of measuring

performance and automation dependence behavior in that environment; and (3) the effect of task complexity on operator performance and the quantification of complexity using Shannon entropy. This study will also seek to provide additional empirical evidence for the dissociation between task load and subjective workload and fatigue ratings.

This study seeks to build on existing work by the Naval Research Laboratory, which is focused on the development of a sensitive suite of performance and user state metrics that can be used for future development, testing, and evaluation of UAV ground control stations; automation; decision support tools; data visualizations; and personnel selection. Finally, this research will investigate all of the above within a highly unique population: Student Naval Aviators (SNA) and Student Naval Flight Officers (SNFOs). SNAs/SNFOs tend to be a highly homogenous and range-restricted population relative to undergraduate university students, who are the usual subjects of such research.

3 Method

3.1 Participants

Participants in this study included 81 Student Naval Aviators (SNAs) and Student Naval Flight Officers (SNFOs) at the Naval Aerospace Medical Institute (NAMI) in Pensacola, Florida. The group included both male ($n = 71$), female ($n = 9$), and unspecified gender ($n = 1$) participants, who averaged 23.8 years in age (range: 21–30 years). Of the participants who reported their visual acuity, most possessed uncorrected or corrected vision of 20/20 or better ($n = 61$), and all but one had corrected or uncorrected vision of 20/40 or better. The remaining participant had 20/50 vision. Ten participants did not report their visual acuity, but all can be assumed to possess visual acuity above the minimum required for an SNFO, which is 20/20 corrected (U.S. Navy, 2018). A small number of participants wore contacts ($n = 2$) or glasses ($n =$

5) during the experiment. The majority of participants were right eye ($n = 65$) and right hand ($n = 75$) dominant, but there was a solid minority of left eye ($n = 16$) and left hand ($n = 6$) dominant participants. Participants spent an average of 17.4 hours a month gaming ($SD = 23.987$), and reported their subjective skill levels as follows: novice ($n = 34$), intermediate ($n = 29$), and expert ($n = 18$). Two participants reported previous commercial or military UAV operational experience. None of the participants currently have, or have ever had, a pilot's license.

3.2 Apparatus

3.2.1. SCOUT.

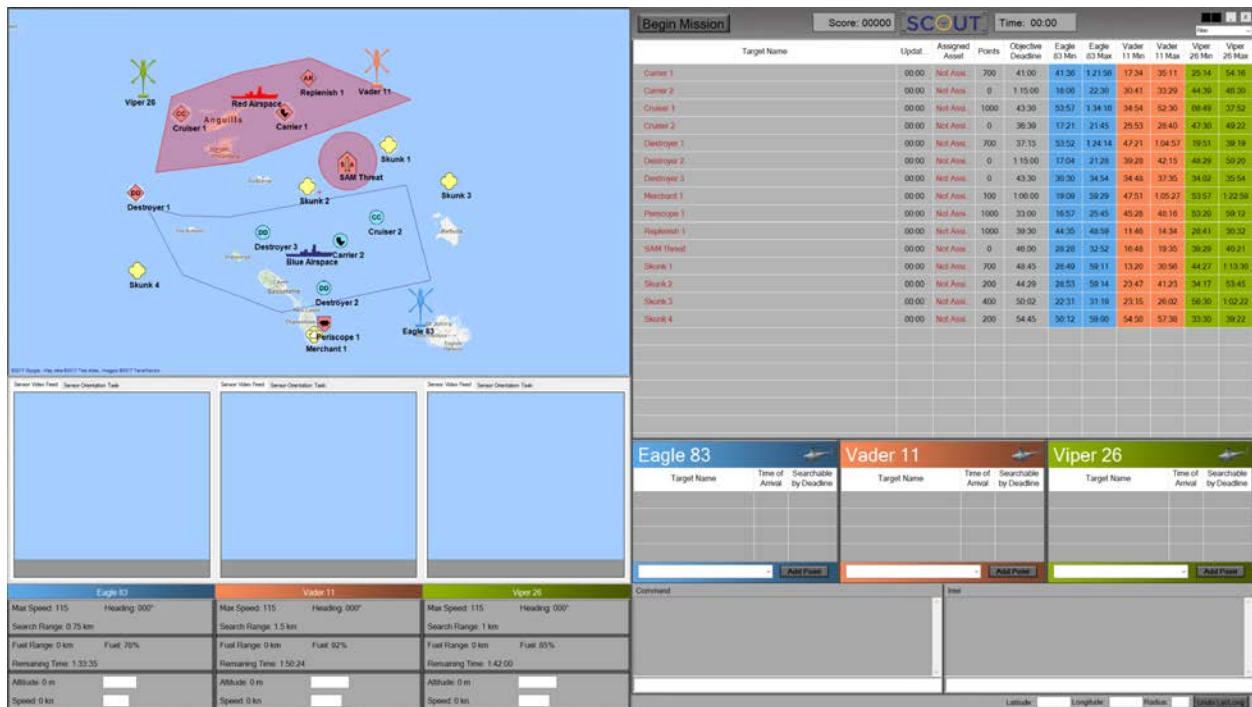


Figure 3.1. The single-screen SCOUT GUI variant.

The Supervisory Control Operations User Testbed (SCOUT), developed by the Naval Research Laboratory (NRL), is a realistic simulation environment for assessing single operator performance monitoring multiple unmanned aerial vehicles (UAVs). It is designed to replicate the complexity, noise, and uncertainty associated with military UAV control. In addition, it includes tasks representative of current operators' primary roles: route planning, airspace management, communication, and monitoring (Coyne & Sibley, 2015b) (Figure 3.1).

During a SCOUT mission, participants manage three heterogeneous helicopter UAVs. To meet mission goals, they must decide how to best allocate the UAVs to locate targets while simultaneously completing several subtasks, including maintaining communication with command and intelligence personnel via chat, updating UAV parameters, and monitoring their sensor feeds and airspace. Points are assigned to various actions based on their mission priority and the goal is to obtain as many points as possible.

The single-screen SCOUT variant was run on 14 custom PC workstations, each equipped with a 25-inch Acer monitor with a display resolution of 2560 x 1440. Participants sat approximately 65 cm. from the display. For a more detailed walkthrough of the SCOUT test bed and its subtasks, please see the continuation of this section in Appendix A.

3.3. Procedure

Table 3.1

Experiment Schedule

TIME (HRS:MIN:SEC)	ACTIVITY
0:15:00	Informed consent, demographic survey, and SCOUT setup and orientation*
0:30:00	Interactive SCOUT training

0:13:45	SCOUT practice scenario
0:48:00	SCOUT experimental scenario
0:03:00	Trust in automation survey (paper version)
0:00:00	Debrief (paper form)
1:49:45 TOTAL	

Note. * = Although instructions were provided on-screen, the experimenter verbally walked participants through the setup tasks.

3.3.1. Overview. The experiment utilized a 2x2x2 mixed MANOVA design. The reliability of the payload task automation was treated as a between-subjects factor. Task load and task complexity were treated as within-subjects factors. Dependent variables included the following subtask performance measures: maximum expected value (per block) on the UAV routing task, accuracy on the payload (i.e., target identification) task, and throughput for information and vehicle/target update requests from Command and Intelligence personnel. Additional dependent variables included participant responses to the Crew Status Survey, a subjective fatigue and workload questionnaire; participant responses to a subjective self-confidence and trust in automation survey; and percent agreement with the payload task automation, which served as an indicator of automation dependence. These dependent variables are described in further detail in section 3.5. The entire experiment, including participant training and debriefing, took just under two hours to complete and followed the schedule outlined in Table 3.1.

3.3.2. Demographic survey. After completing all necessary informed consent and data release documentation, participants completed a short demographic survey. The survey requested participants to report the following: age, gender, hours per month spent playing video/computer games, gaming skill level (novice, intermediate, or expert), whether they wore glasses or contacts (during the experiment), dominant hand, dominant eye, visual acuity, commercial/military UAV operational experience, and whether they had a pilot's license.

3.3.3. SCOUT training. Each participant completed approximately 30 minutes of interactive training on the operation of SCOUT. The training was fully automated and self-paced, though written instructions indicated that participants should aim to complete the training in about 30 minutes. The training included a text-based walkthrough of the test bed with guided practice and covered such topics as UAV capabilities, route planning, communication, airspace monitoring, and target searching.

3.3.4. SCOUT practice scenario. After completing the 30-minute training course, participants completed a 13:35 minute practice scenario designed to ensure adequate baseline knowledge of SCOUT. During the training period and practice scenario, participants were encouraged to ask the experimenter for clarification on any aspect of the test bed or its operation that they found unclear.

3.3.5. The SCOUT experimental scenario. Once participants finished the practice scenario and asked the experimenter for additional clarification on the SCOUT controls (if needed), they proceeded to an approximately 48-minute SCOUT experimental scenario. The experimental scenario consisted of an untimed planning period (which typically took participants up to 10 minutes) followed by a 34:15 minute mission with five 45-second workload/fatigue freeze probes. During the planning period, the participant formulated the best plan for sending their three helicopter UAVs to five possible target areas. The goal of each scenario was to obtain as many points as possible.

Table 3.2

Point Values Associated with Various User Actions

Action	Points
Located target	Variable (0 – 3000)
Answered Command or Intel information request	25

Made requested UAV/target status update	25
Updated UAV directional sensor	100
Reported UAV position at 5 min. \pm 1 min. to target arrival	100
Reported UAV position at 5 min. \pm 2 min. to target arrival	50
Selected potential target	25
Missed potential target	0
Selected distractor target	-25
Incurred ROZ without approval	Variable (-3000 – 0)

Once the scenario began, in order to score points, participants responded to text communications from command and intelligence personnel, made updates to UAV and target parameters on request, reported UAV position when five minutes out from a target search area and—for the largest point gains—located targets by their deadlines. To successfully locate a target, participants had to monitor the sensor feed of the searching UAV once it arrived over the target search area. The Crew Status Survey, a brief workload and fatigue questionnaire, was administered during five task freezes but did not offer a point reward (Ames & George, 1993).

New targets appeared over the course of each scenario, which meant participants had to continually re-plan if they wished to maximize their points. However, participants had to simultaneously monitor their airspace since points were lost if they incurred a Restricted Operating Zone (ROZ) without permission. To avoid losing points, participants were required to request access prior to crossing a ROZ boundary. See Table 3.2 for a summary of the point values associated with different user actions.

3.4 Independent Variables

3.4.1. Automation reliability. At the beginning of the experiment, participants were randomly sorted into a low automation reliability group ($n= 40$) and a high automation reliability group ($n= 41$). As previously stated, SCOUT included a sensor monitoring component. Once a

UAV arrived at a target, participants had to monitor that UAV's sensor feed in order to locate the target, which was surrounded by distractors. To assist participants with locating targets, SCOUT included automation that selected potential target matches; selected targets were enclosed in a brown box. In cases of automation error, the participant could manually deselect the target before reached the bottom of the sensor feed to avoid losing points. Participants received 25 points for each correctly identified potential target, but lost 25 points for each erroneously selected distractor target. Missed potential targets did not result in point loss.

Table 3.3

Hits, Misses, False Alarms, and Correct Rejections for Low and High Automation Reliability Conditions

Automation Reliability	Hit	Miss	False Alarm	Correct Rejection
High (97%)	97.0%	3.00%	3.00%	97.0%
Low (92.5%)	100%	0.00%	15.0%	85.0%

In the high automation reliability condition, the automation was 97% reliable and was capable of effectively discriminating between potential targets and distractors. More specifically, the automation had a 97.0% hit rate (correct selection of a potential target), a 3.00% miss rate (failure to select a potential target), a 3.00% false alarm rate (erroneous selection of a distractor), and a 97.0% correct rejection rate (correct dismissal of a distractor).

In the low automation reliability condition, the automation had a liberal response criterion and was thus false-alarm prone. The automation, which was 92.5% reliable, required the participant to manually deselect numerous false alarms (i.e., erroneously selected distractor targets) to avoid point loss. More specifically, the automation had a 100% hit rate (correct selection of a potential target), a 0.00% miss rate (failure to select a potential target), a 15.0% false alarm rate (incorrect selection of a distractor), and an 85.0% correct rejection rate (correct dismissal of a distractor) (Table 3.3). Since erroneously selected false alarms were indicated by

an auditory alert, this condition resulted in excessive alerts and potential participant annoyance, distraction, and disruption of performance.

Table 3.4

Experimental Scenario Timeline

Scenario Variant	Block 1 (0:00–8:33)	Block 2 (8:34–17:07)	Block 3 (17:08–25:41)	Block 4 (25:42–34:15)
Alpha*/Echo†	Low C/Low TL	Low C/High TL	High C/Low TL	High C/High TL
Bravo*/Foxtrot†	Low C/High TL	Low C/Low TL	High C/High TL	High C/Low TL
Charlie*/Golf†	High C/Low TL	High C/High TL	Low C/Low TL	Low C/High TL
Delta*/Hotel†	High C/High TL	High C/Low TL	Low C/High TL	Low C/Low TL
FWP1	FWP2	FWP3	FWP4	FWP5

Note: * = high automation reliability; † = low automation reliability; C = complexity; TL = task load; FWP = Fatigue and Workload Probe/CSS. Times listed for each block represent scenario clock time and thus do not include the time allotted for task freezes.

Within the low and high automation reliability groups, task load and task complexity were varied between low and high levels during the experimental scenario. The experimental scenario was pre-scripted in eight versions, using a Latin Square design, to account for order effects; while the rates at which targets and messages appeared differed over time between scenario variants, they were otherwise as identical as possible (Table 3.4).

3.4.2. Task load.

Table 3.5

Message and Target Frequency for Low and High Task Load Conditions

TASK	MESSAGES (MIN. AND SEC.)			TARGETS (MIN. AND SEC.)		
	Number	Frequency	±	Number	Frequency	±
LOW	3	2:51	0:34	1	8:33	1:43
HIGH	32	0:16	0:03	4	2:08	0:26

Note. The frequency is an average. Actual times varied up to 20%.

Operator task load was manipulated by increasing or decreasing both the frequency of new targets and the frequency of messages from Command and Intelligence. In a low task load block, participants received chat messages from Command or Intelligence every 2:51 minutes, on average. In addition, only one new target appeared on the moving map display (approximately midway through the block). In a high task load block, participants received chat messages, on average, every 16 seconds. Four new targets appeared on the moving map display (every two minutes and eight seconds, on average). All times varied, at random, up to 20% to avoid participant detection of a pattern. See Table 3.5 for a summary of these times.

3.4.3. Task complexity.

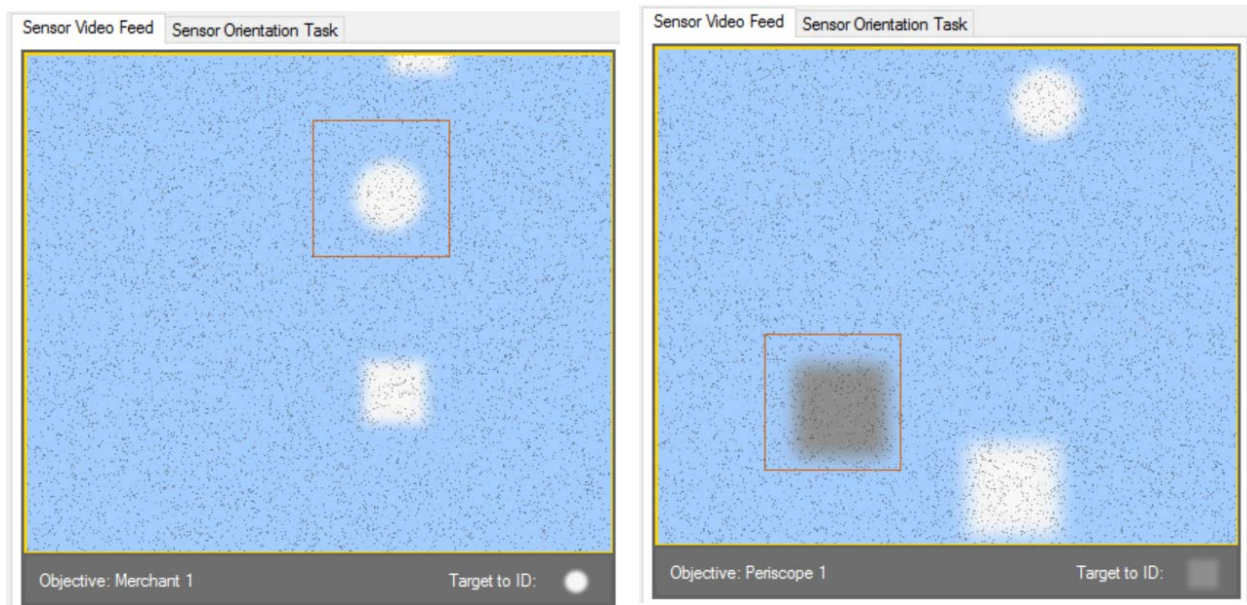


Figure 3.2. Sample low and high complexity sensor pictures. In the high complexity condition (right), large targets are 30% larger than small targets.

In low task complexity trials, participants were required to identify a target amidst distractor targets that differed from it on one dimension: shape (circle or square). For example, participants could be prompted to locate Merchant 1, a circle amongst square distractors. In a high task complexity trial, participants were required to identify a target surrounded by distractors that differed from it on three dimensions: shape (circle or square), color (gray or white), and size (small and large). For example, participants could be asked to locate Periscope 1, a large, gray square. Figure 3.2 shows an example of low and high complexity sensor pictures.

The precise difference in task complexity between the low and high payload complexity conditions was operationalized using Shannon entropy (Teixeira, Matos, Souto, & Antunes, 2011). As previously discussed, the payload task involved the discrimination of three target features, each with two possible values:

$$\text{Shape (A)} = \{\text{circle, square}\} = \{a, a'\}$$

$$\text{Color (B)} = \{\text{gray, white}\} = \{b, b'\}$$

$$\text{Size (C)} = \{\text{big, small}\} = \{c, c'\}$$

Therefore, there were eight possible states of the target:

$$x_1 = abc$$

$$x_2 = abc'$$

$$x_3 = ab'c$$

$$x_4 = ab'c'$$

$$x_5 = a'bc$$

$$x_6 = a'bc'$$

$$x_7 = a'b'c$$

$$x_8 = a'b'c'$$

Table 3.6
Target Feature Probabilities for the Low Complexity Payload Task

	$P(x_i)$
$x_1 = abc$	0
$x_2 = abc'$	0
$x_3 = ab'c$	0
$x_4 = ab'c'$	0.5
$x_5 = a'bc$	0
$x_6 = a'bc'$	0
$x_7 = a'b'c$	0
$x_8 = a'b'c'$	0.5

Table 3.7
Target Feature Probabilities for the High Complexity Payload Task

	$P(x_i)$
$x_1 = abc$	0.125
$x_2 = abc'$	0.125
$x_3 = ab'c$	0.125
$x_4 = ab'c'$	0.125
$x_5 = a'bc$	0.125
$x_6 = a'bc'$	0.125
$x_7 = a'b'c$	0.125
$x_8 = a'b'c'$	0.125

Table 3.6 and Table 3.7 display the probability of occurrence of targets with each combination of features in the low and high complexity payload task, respectively. The probabilities assume that the target features have an equal probability of appearing on the payload display, which is indeed the case.

Thus, for the low complexity payload task, the information entropy is

$$H(X) = -\sum P(x_i) \log_2 P(x_i) = -0.5 * \log_2(0.5) * 2 = 1 \quad (3.1)$$

For the high complexity payload task, the information entropy is

$$H(X) = -\sum P(x_i) \log_2 P(x_i) = -0.125 * \log_2(0.125) * 8 = 3 \quad (3.2)$$

3.5. Dependent Variables

3.5.1. Performance. Performance measures included maximum expected value per block on the UAV routing task, accuracy on the payload task, and throughput for information and vehicle/target update requests from Command and Intelligence. Ideally, throughput would also

be used to assess performance on the payload task but, due to a data-logging problem, reaction time for the task is unavailable. However, since there is significant range restriction on the reaction time of the payload task because targets only appear on the feed for seven seconds (in contrast to the full minute participants have to respond to Command and Intelligence), the impact of the reaction time data loss should be minimal. Points, although displayed to the operator for motivational purposes, is a composite measure that takes into consideration all the previous factors, and is thus not included in the analysis.

3.5.1.1. UAV routing task. Performance on the UAV routing task is defined as an operator's adjusted expected value, which is calculated by block using the following formula:

$$EV_{adjusted} = \frac{EV_{maxp}}{EV_{maxg}} = \frac{\Sigma[(point\ value_p)(searchable_p)]}{\Sigma[(point\ value_g)(searchable_g)]} \quad (3.3)$$

EV_{maxp} = maximum expected value per block for a given participant, or the sum of the products of each assigned target's point value and the percentage of the target searchable by its deadline

EV_{maxg} = the maximum expected value per block that was achieved by any participant (i.e., the participant with the highest EV for that block)

3.5.1.2. Chat communication task. Performance on the communication task is defined by operators' throughput. Throughput is a composite measure equal to the number of correct task responses (e.g., the number of correctly answered chat messages) divided by the cumulative reaction times, both correct and incorrect (Thorne, 2006).

3.5.1.3. Payload task. Performance on the payload task is defined by percent accuracy, according to the following formula:

$$Accuracy = \left(\frac{Hits_p + CR_p}{Hits_b + CR_b} \right) 100 \quad (3.4)$$

p = participant selections (including agreement with automated selections and overrides)
b = baseline (selections made by the automation prior to human operator interference)

According to this formula, accuracy is the sum of the hits and correct rejections made by a human operator working in conjunction with the automation divided by the sum of the hits and correct rejections made by the automation prior to human operator interference. Higher accuracy scores indicate better performance. An accuracy score of 100% indicates that the participant exhibited the same accuracy as the automation if it were left to perform the task without any user interference. An accuracy score greater than 100% indicates that the participant performed better than the automation. In other words, they overrode the automation on at least one occasion to improve its performance. The greater their accuracy score over 100%, the more automation errors the participant successfully caught and corrected. Conversely, if a participant's accuracy score falls below 100%, it means they performed worse than if they had left the automation to handle the task in isolation and they erroneously deselected targets that were correctly identified by the automation.

3.5.2. Automation dependence. An operators' dependence on the payload task automation was operationalized as their percent agreement with the automation, or the percentage of their responses that followed the automation's recommendation. A percent agreement score under 100% indicated that a participant did not always follow the recommendation of the automation and manually selected and/or deselected potential targets previously identified by the automation before it could execute its decision. A score of 100% indicated that a participant always followed the automation.

3.5.3. Subjective measures of operator state.

3.5.2.1. Crew status survey (CSS). The Crew Status Survey was used to assess operator fatigue and workload (Ames & George, 1993). Participants rated their current level of fatigue on a seven-point scale that ranged from one (fully alert) to seven (completely exhausted). Next, participants rated both the average

and maximum level of workload they experienced during the past work period (since the beginning of the scenario or the last questionnaire, whichever came last) on a seven-point that ranged from one (nothing to do) to seven (overloaded).

3.5.2.2. Trust in automation survey. A brief four-question Likert-type survey was administered upon completion of the SCOUT experimental scenario. The survey, modeled after Lee and Moray's validated (1994) scale, is used to rate an operator's trust in an automated system and their self-confidence that they could perform the same task manually. The purpose of this survey was to gauge whether participants in the low and high automation reliability groups noticed a difference in the reliability of the automated target selection and if their trust and self-confidence were accordingly impacted.

4 Results

4.1. Overview

The purpose of this study was to assess the effect of task load, environment complexity, and automation reliability on UAV operators' performance, subjective workload and fatigue, automation dependence, and trust in the automation.

The data analysis aims to support answers to the following questions:

- Do differences in task load, environment complexity, and automation reliability affect participants' task performance? Task performance is defined as adjusted expected value on the UAV routing task, accuracy on the automated payload task, and throughput on the chat communication task.
 - Accuracy was used in place of throughput to assess performance on the automated payload task. Due to data loss, response times were unavailable and throughput could not be calculated. However, the impact on results was minimal because

response time was range restricted on the payload task; participants only had seven seconds to react to each target.

- Do differences in task load, environment complexity, and automation reliability affect participants' subjective workload?
- Do differences in task load, environment complexity, and automation reliability affect participants' subjective fatigue?
- Do differences in task load, environment complexity, and automation reliability affect operators' automation dependence?
- Do differences in automation reliability affect participants' subjective ratings of trust in the automation? (I.e., do participants notice a difference in the reliability of the automation?)

With the exception of the analysis of the subjective trust rating data, which was conducted using t-tests, all analyses were performed using either a mixed analysis of variance (ANOVA) or a linear mixed-model (LMM) approach. The mixed ANOVA was employed in all cases where less than 5% of the data were missing because it is ubiquitous, easy to use, and, most importantly, appropriate for the study design. However, an LMM approach was used instead of ANOVA when more than 5% of the data were missing because mixed-effects models are much more robust against missing data than ANOVA as long as the data are missing at random (MAR) (Gueorguieva & Krystal, 2004). In all cases where an ANOVA was used, an equivalent LMM analysis was run on the same data set as a precaution. Although the results of these redundant LMM analyses are not reported herein, the results were very similar to the results obtained with the ANOVA and the significance and/or non-significance of the variables did not change for any of the analyses.

4.2 UAV Operator Performance

4.2.1. Routing task performance (adjusted expected value). Participant performance on the UAV routing task was assessed using their adjusted expected value, or the sum of the products of the point values and searchable by deadline percentages of all targets a participant assigns to their three UAVs during each experimental block divided by the maximum expected value per block that was achieved by the participant with the highest EV for that block.

The Little's MCAR test for these variables resulted in $\chi^2 = 29.129$ ($df = 17$; $p = 0.033$), which indicates the data is missing not at random (MNAR). Further examination of the data revealed that more data was missing for the high task load blocks (8.6–11.1%) than the low task load blocks (1.2–6.2%). None of the 81 participants included in the analysis dropped out of the study, so these “missing” values are not truly missing and, in fact, indicate that participants simply did not assign any new targets to their UAVs during that block. It is possible that participants, when task saturated, did not notice new targets appearing on the moving map or chose not to spend time assigning them to UAVs and, thus, received an expected score of zero for that block. It is also possible that non-optimal decision-making during the prior block resulted in their UAVs being out of range of new targets. If that were the case, participants might have been reluctant to expend the time to assign futile targets to their UAVs when they could otherwise attend to alternate tasks to gain points. Since an identifiable pattern exists in the missing data, the data are missing not at random (MNAR) and the results of this analysis should be interpreted with a degree of caution.

Adjusted expected value scores were not normally distributed, as assessed by the Kolmogorov-Smirnov test ($p < 0.05$). However, since values for skewness and kurtosis between -

2 and +2 are acceptable to demonstrate normal univariate distribution (George & Mallery, 2009), the analysis was continued without data transformation for easier interpretation.

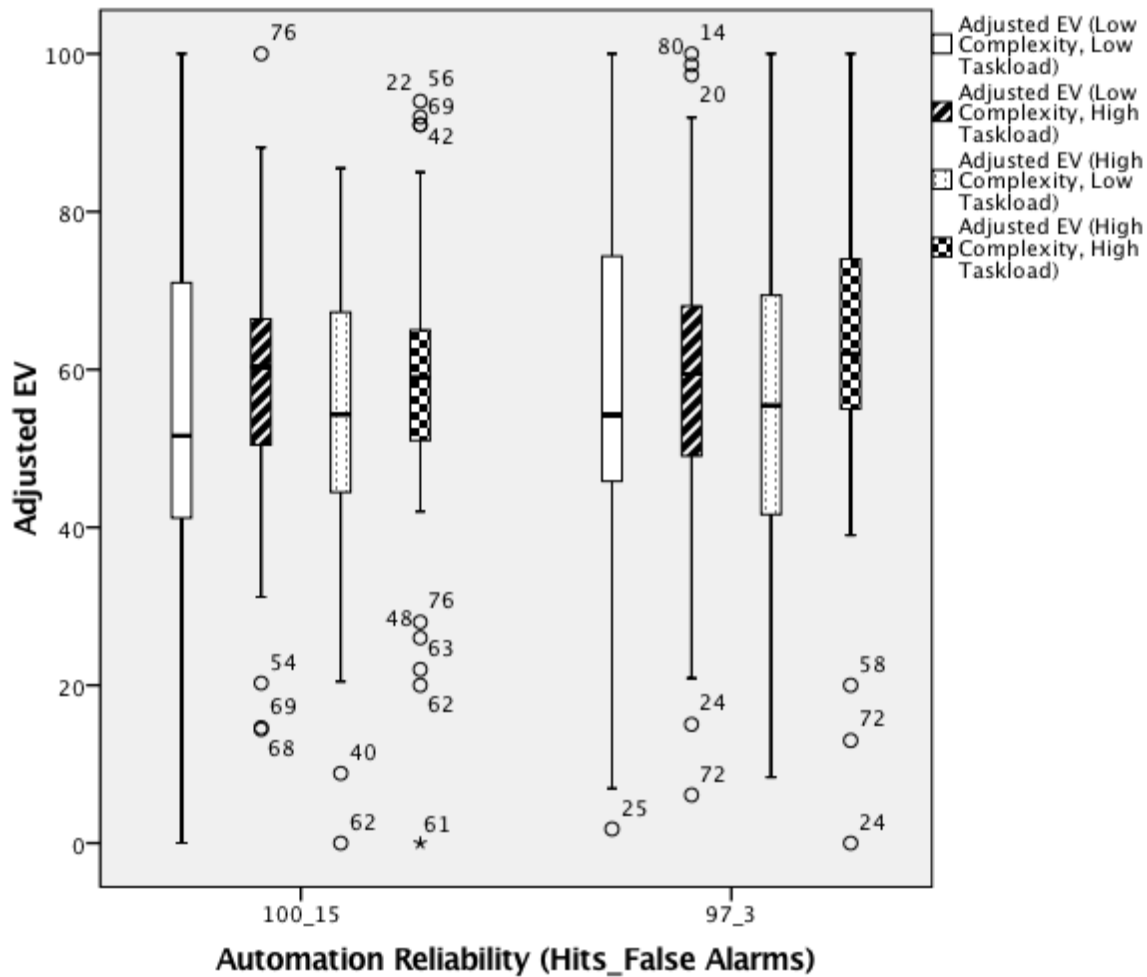


Figure 4.1. Adjusted expected values by experimental block. Adjusted expected values for the unreliable automation group (100% hit rate, 15% false alarm rate) are displayed on the left and adjusted expected values for the reliable automation group (97% hit rate, 3% false alarm rate) are displayed on the right.

There were 24 outliers in the data, as assessed by inspection of a boxplot (Figure 4.1). The outliers were kept in the analysis because they are genuinely unusual values and not the result of measurement or data entry errors. The outliers do not have an appreciable effect on the analysis as assessed by a comparison of the results with and without the outliers.

A restricted maximum likelihood linear mixed-model (REML LMM) analysis was run on the adjusted expected value data. The LMM approach was used instead of a mixed ANOVA because LMM is able to accommodate missing data and simply imputing the missing data to run the GLM is not appropriate. Expectation-Maximization (EM) imputation is not ideal when more than 5% of the data is missing, particularly when there is a pattern to the missing data, as is the case here (Schafer, 1999; Tabachnick and Fidell, 2013). Multiple imputation, in theory, could be used to impute the missing data prior to running a GLM analysis, but there is no agreed upon method in the literature to pool multiply imputed datasets for a 2 x 2 x 2 mixed ANOVA. Thus, a LMM analysis is the most appropriate choice for this dataset.

$$\text{Adjusted EV} \sim \text{Task Load} + \text{Task Complexity} + \text{Automation Reliability} + (1|\text{Participant}) + (1|\text{Block}) + \varepsilon \quad (4.1)$$

As previously stated, an REML LMM analysis was run to assess adjusted expected value as a function of task load, task complexity, and automation reliability. Task load, task complexity, and automation reliability (and their two-way interaction terms) were entered into the model as fixed factors. Fixed factors are those where all levels of interest are controlled for by the study design. Participant ID was included in the model as a random factor (i.e., a “grouping variable”) to resolve the violation of non-independence of observations by assuming a different “baseline” adjusted expected value for each participant since repeated measures in a mixed design are not independent. Experimental block was added as an additional random factor to account for by-block variation in adjusted expected values, which are also not independent (Magezi, 2015; Winter, 2013). This information is summarized in Equation 4.1, which also includes a general error term “ ε .”

There was homogeneity of variances, as assessed by Levene's test for equality of variances ($p > 0.05$).

Multiple models (compound symmetry; first order autoregressive, AR(1); and unstructured) were fit, and one was ultimately selected, via a penalized likelihood approach. More specifically, the BIC (Bayesian information criterion) and AIC (Akaike information criterion) were compared for various models and the model with the lowest BIC and AIC values was selected. Since Type I error control was considered a higher priority than loss of power, the AIC value was given more weight when the BIC and AIC diverged. Selecting an overly simplified model inflates the Type I error rate and lower BIC values tend to be associated with less complex models. Conversely, overly complex models, which are penalized by BIC and characterized by higher BIC values, result in loss of power (Guerin & Stroup, 2000; Seltman, 2018).

The selected model utilized an unstructured covariance structure (Gurka, Edwards, & Muller, 2011). Mauchly's test of sphericity was not assessed because sphericity or compound symmetry was not assumed in the model. However, even if sphericity was assumed, there are only two levels of both within-subjects factors. Therefore, there would only be one paired difference for each and the assumption of sphericity would automatically be met.

None of the two-way interactions for route task performance were significant. First, there was no statistically significant simple two-way interaction between complexity and reliability, $F(1, 75.795) = 0.005, p = 0.946, \text{partial } \eta^2 = 0.000$. Second, there was also no statistically significant simple two-way interaction between task load and reliability, $F(1, 70.186) = 0.007, p = 0.932, \text{partial } \eta^2 = 0.000$. Finally, there was no statistically significant simple two-way interaction between complexity and task load, $F(1, 78.819) = 0.003, p = 0.356, \text{partial } \eta^2 = 0.011$.

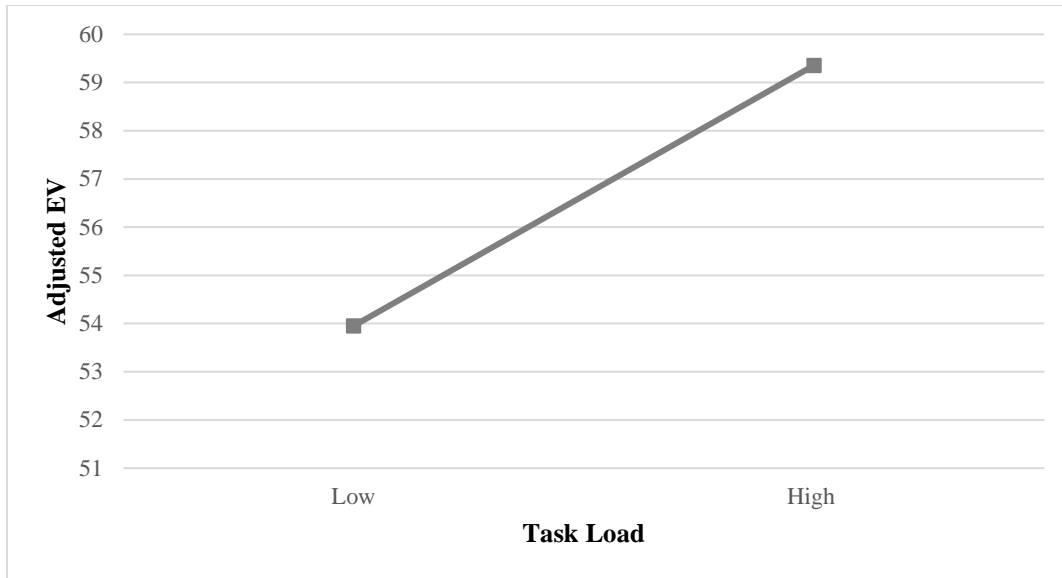


Figure 4.2. Main effect of task load on adjusted expected value on the UAV routing task.

However, there was a statistically significant main effect of task load, $F(1, 69.586) = 9.688, p = 0.003, \text{partial } \eta^2 = 0.1222$ (Figure 4.2). As participant task load increased, their performance on the UAV routing task increased. Participants' adjusted expected values for the UAV routing task were significantly better in the high task load condition ($M = 59.39, SD = 19.89$) than the low task load condition ($M = 53.97, SD = 21.58$). The main effects of payload task complexity, $F(1, 75.531) = 0.053, p = 0.818, \text{partial } \eta^2 = 0.001$, and payload task automation reliability, $F(1, 74.059) = 1.428, p = 0.236, \text{partial } \eta^2 = 0.019$, on participants' adjusted expected values for the concurrent UAV routing task were not significant.

4.2.2. Payload task performance (accuracy). Participant performance on the payload task was assessed using their percent accuracy, or the sum of the hits and correct rejections made by the participant working in conjunction with the automation divided by the sum of the hits and correct rejections made by the automation prior to participant interference. Higher accuracy scores indicate better performance.

The Little's MCAR test obtained for these variables resulted in $\chi^2 = 29.275 (df = 17; p = 0.032)$, which indicates the data are MNAR. Further examination of the data revealed that

substantially more data are missing for the low task load blocks (13.6–22.2%) than the high task load blocks (1.2–3.7%). This pattern most likely exists because fewer new targets appeared during the low task load blocks. Since there are fewer opportunities to assign targets to UAVs, and thus fewer active target searches that could take place during a low task load block, it is more likely that poor planning could result in no target searches occurring at all during that period.

Participants' percent accuracy was not normally distributed, as assessed by the Kolmogorov-Smirnov test ($p < 0.05$). A square-root transformation was applied to the data to correct a moderate negative skew (Tabachnick & Fidell, 2013). However, the accuracy score distributions for the low and high task load conditions remained slightly leptokurtic (kurtosis values of 2.049 and 3.195, respectively). All other skewness and kurtosis values fell between the -2 and +2 range acceptable to demonstrate normal univariate distribution after transformation (George & Mallery, 2009).

analysis of the adjusted expected value data above, the LMM approach was used instead of a mixed ANOVA because LMM is able to accommodate missing data.

$$\text{Percent Accuracy} \sim \text{Task Load} + \text{Task Complexity} + \text{Automation Reliability} + (1|\text{Participant}) + (1|\text{Block}) + \varepsilon \quad (4.2)$$

Task load, task complexity, and automation reliability (and their two-way interaction terms) were entered into the model as fixed factors. Participant ID and experimental block were included in the model as random factors (Magezi, 2015; Winter, 2013). This information is summarized in Equation 4.2, which also includes the general error term “ ε .”

There was homogeneity of variances, as assessed by Levene's test for equality of variances ($p > 0.05$).

Multiple models (compound symmetry; first order autoregressive, AR(1); and unstructured) were fit via a penalized likelihood approach or, more specifically, through comparison of their BIC and AIC values (Seltman, 2018). Although the AR(1) covariance structure produced a smaller AIC value, the selected model utilized an unstructured covariance structure since its BIC value was smaller and minimization of potential Type I error was desired (Guerin & Stroup, 2000; Gurka, Edwards, & Muller, 2011). Mauchly's test of sphericity was not assessed because sphericity or compound symmetry was not assumed in the model. However, even if sphericity was assumed, there are only two levels of both within-subjects factors. Therefore, there would only be one paired difference for each and the assumption of sphericity would automatically be met.

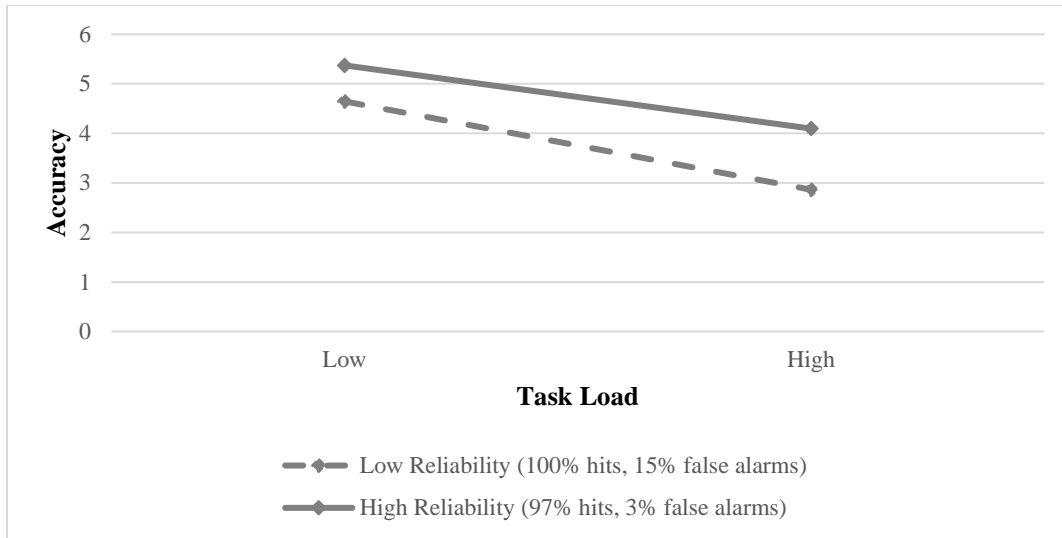


Figure 4.4. Reliability and task load interaction for payload task accuracy. This data was square-root transformed.

There was a statistically significant simple two-way interaction between reliability and task load, $F(1, 71.776) = 7.804, p = 0.007, \text{partial } \eta^2 = 0.098$ (Figure 4.4). However, there were no statistically significant simple two-way interactions between reliability and complexity, $F(1, 70.828) = 0.015, p = 0.903, \text{partial } \eta^2 = 0.000$, or complexity and task load, $F(1, 72.783) = 0.897, p = 0.347, \text{partial } \eta^2 = 0.012$.

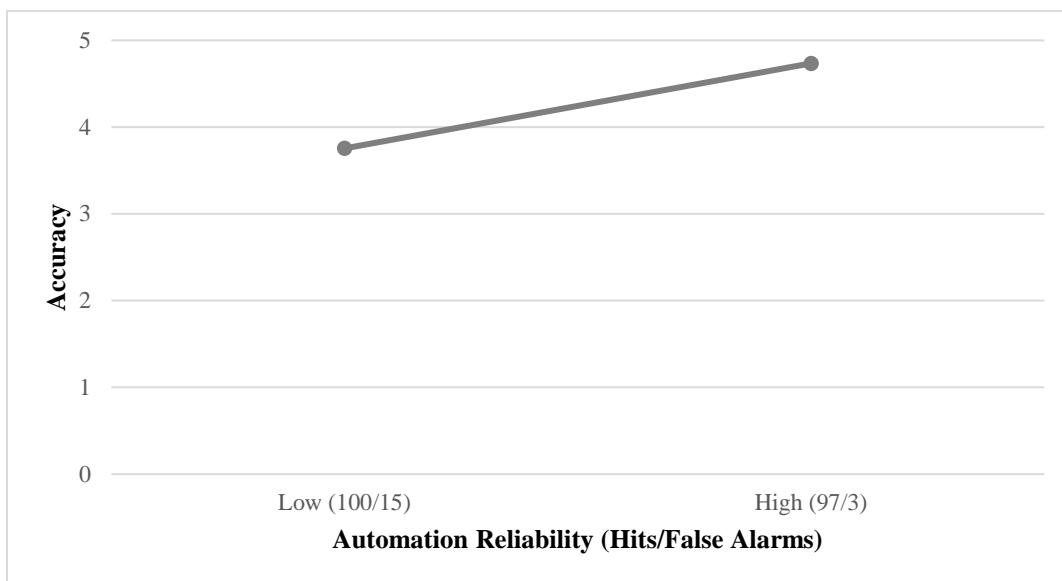


Figure 4.5. Main effect of automation reliability on payload task accuracy. This data was square-root transformed.

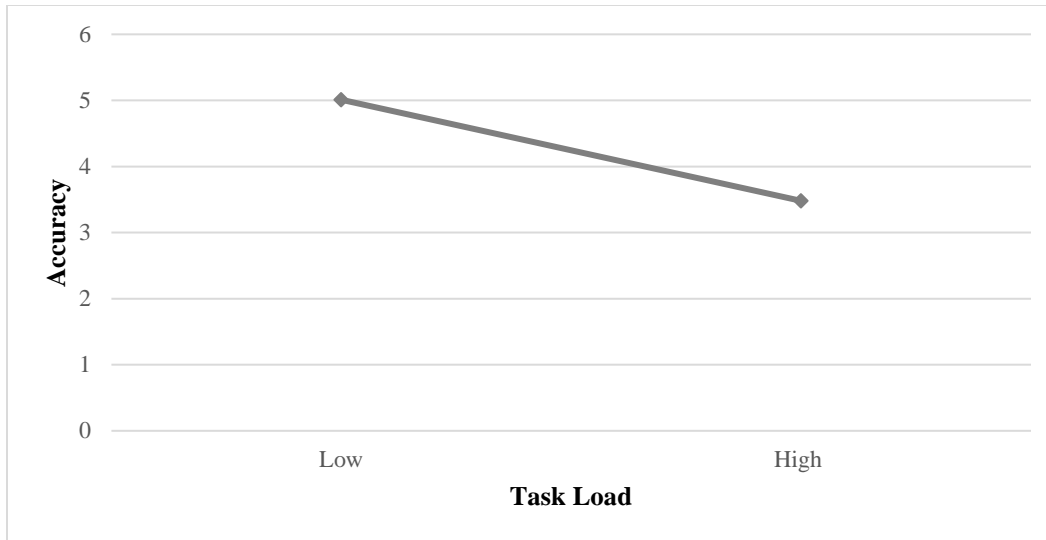


Figure 4.6. Main effect of task load on payload task accuracy. This data was square-root transformed.

The main effects of automation reliability, $F(1, 74.407) = 13.782, p = 0.000$, partial $\eta^2 = 0.156$ (Figure 4.5), and task load, $F(1, 71.821) = 9.688, p = 0.000$, partial $\eta^2 = 0.799$ (Figure 4.6), were also significant. There was no significant main effect for payload task complexity, $F(1, 64.665) = 0.342, p = 0.561$, partial $\eta^2 = 0.005$.

Participant's percent accuracy on the payload task generally improved when aided by reliable automation ($M = 4.65, SD = 1.539$) and worsened when aided by unreliable automation with a liberal response criterion ($M = 3.66, SD = 1.671$). In addition, their percent accuracy was inversely related to their task load; their percent accuracy was higher when their task load was low ($M = 4.94, SD = 1.450$), and lower when their task load was high ($M = 3.49, SD = 1.573$).

Moreover, there was a significant interaction between task load and automation reliability. Unreliable automation caused a more marked performance decrement when participant task load increased from low ($M = 4.646, SE = 0.195$) to high ($M = 2.863, SE = 0.201$). This decrease in payload task accuracy due to an increase in task saturation from low ($M = 5.373, SE = 0.194$) to high ($M = 4.097, SE = 0.199$) was partially mitigated by more reliable automation.

4.2.3. Communication task performance (throughput). Participant performance on the chat communication task was operationalized as their throughput. Throughput is a composite measure equal to the number of correctly answered chat messages divided by the cumulative reaction time of all responses, both correct and incorrect (Thorne, 2006).

The Little's MCAR test obtained for these variables resulted in $\chi^2 = 6.425$ ($df = 3$; $p = 0.093$), which indicated the data were missing completely at random (MCAR). Since less than 5% of the data were MCAR, the missing values were imputed using EM.

A 2x2x2 mixed ANOVA was run to understand the effects of task load, payload task complexity, and payload task automation reliability on participants' throughput for the chat communications task. Throughput values were not normally distributed, as assessed by the Kolmogorov-Smirnov test ($p < 0.05$). However, since values for skewness and kurtosis between -2 and +2 are acceptable to demonstrate normal univariate distribution (George & Mallery, 2010), the analysis was continued without data transformation for easier interpretation.

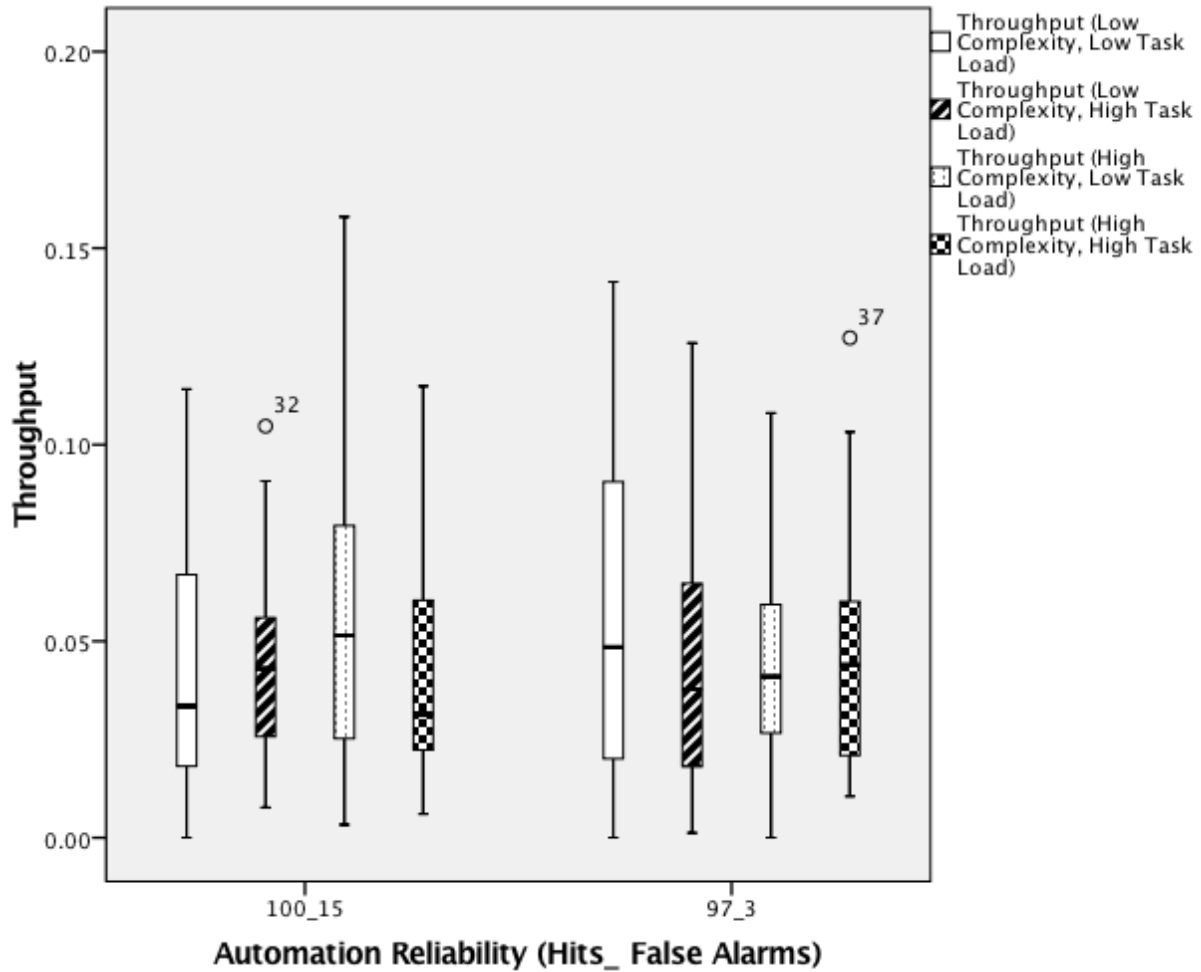


Figure 4.7. Chat communications throughput by experimental block. Throughput values for the unreliable automation group (100% hit rate, 15% false alarm rate) are displayed on the left and throughput values for the reliable automation group (97% hit rate, 3% false alarm rate) are displayed on the right.

There were two outliers in the data, as assessed by inspection of a boxplot (Figure 4.7). The outliers were kept in the analysis because they were genuinely unusual values and not the result of measurement or data entry errors. The outliers did not have an appreciable effect on the analysis as assessed by a comparison of the results with and without the outliers.

Levene's test for equality of variances was significant ($p < 0.05$), indicating that the assumption of homogeneity of variances was violated. Mauchly's test of sphericity was not

assessed because there were only two levels of both within-subjects factors. Therefore, there was only one paired difference for each and the assumption of sphericity was automatically met.

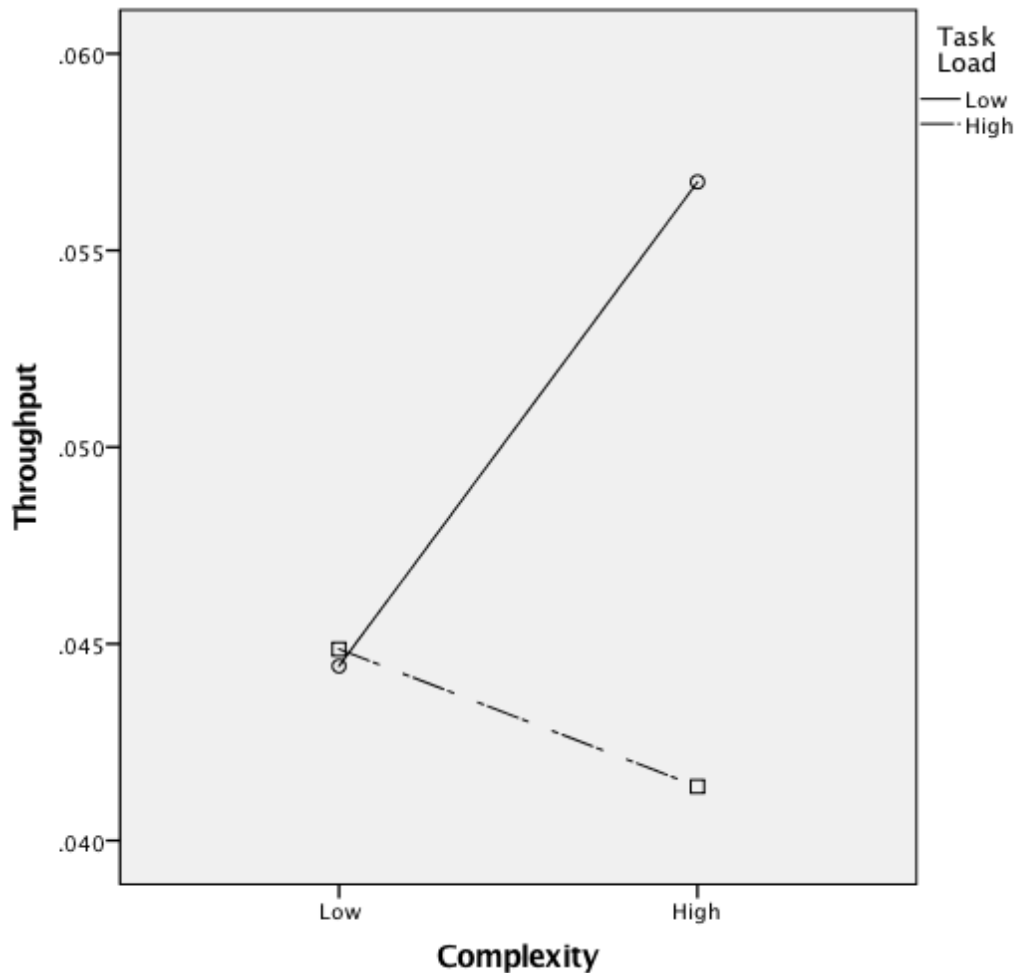


Figure 4.8. Effect of task complexity on communication task throughput values of participants working with unreliable payload task automation.

There was a statistically significant three-way interaction between reliability, task load, and complexity, $F(1, 79) = 4.832, p = 0.031, \text{partial } \eta^2 = 0.058$. Participants working in conjunction with unreliable payload task automation performed similarly on the chat communication task, regardless of overall task load, when the complexity of the payload task was low, although participants given a high task load ($M = 0.045, SE = 0.004$) performed slightly better on the communication task than those given a low task load ($M = 0.044, SE = 0.006$).

When the complexity of the payload task increased, participants relying on unreliable automation during a period of high task load experienced a performance drop on the communication task ($M = 0.041$, $SE = 0.004$). On the other hand, when task load was low, their performance actually improved on the communication task when the payload task became more complex ($M = 0.057$, $SE = 0.005$). Overall, when the payload task automation was less reliable, the difference in performance on the communication task was much more pronounced when the complexity of the payload task was high (Figure 4.8).

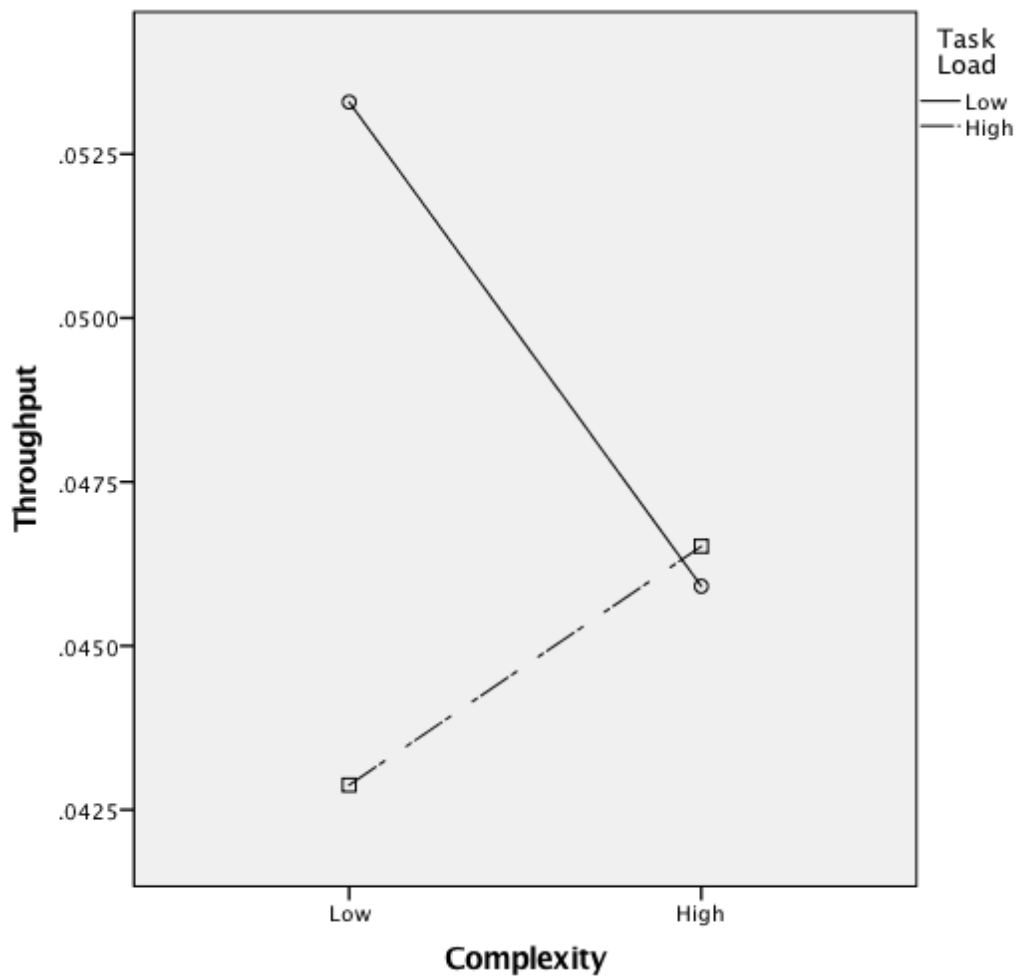


Figure 4.9. Effect of task complexity on communication task throughput values of participants working with reliable payload task automation.

Conversely, participants working with reliable payload task automation performed similarly on the communication task, regardless of task load, when the payload task was more complex, although participants given a high task load ($M = 0.047$, $SE = 0.004$) performed slightly better than those given a low task load ($M = 0.046$, $SE = 0.005$).

When the complexity of the payload task decreased, participants working in conjunction with reliable automation experienced a performance drop on the communication task when task load was high ($M = 0.043$, $SE = 0.004$). On the other hand, when task load was low, their communication performance improved as the payload task became less complex ($M = 0.053$, $SE = 0.006$). Overall, when the payload task automation was more reliable, the difference in performance on the communication task was much more pronounced when the complexity of the payload task was low (Figure 4.9). However, given the relative scarcity of chat messages in the low condition, this three-way interaction should be interpreted with caution.

There was no statistically significant simple two-way interaction between complexity and reliability, $F(1, 79) = 0.592$, $p = 0.444$, partial $\eta^2 = 0.007$. There was no statistically significant simple two-way interaction between task load and reliability, $F(1, 79) = 0.296$, $p = 0.588$, partial $\eta^2 = 0.004$. There was no statistically significant simple two-way interaction between complexity and task load, $F(1, 79) = 0.153$, $p = 0.697$, partial $\eta^2 = 0.002$.

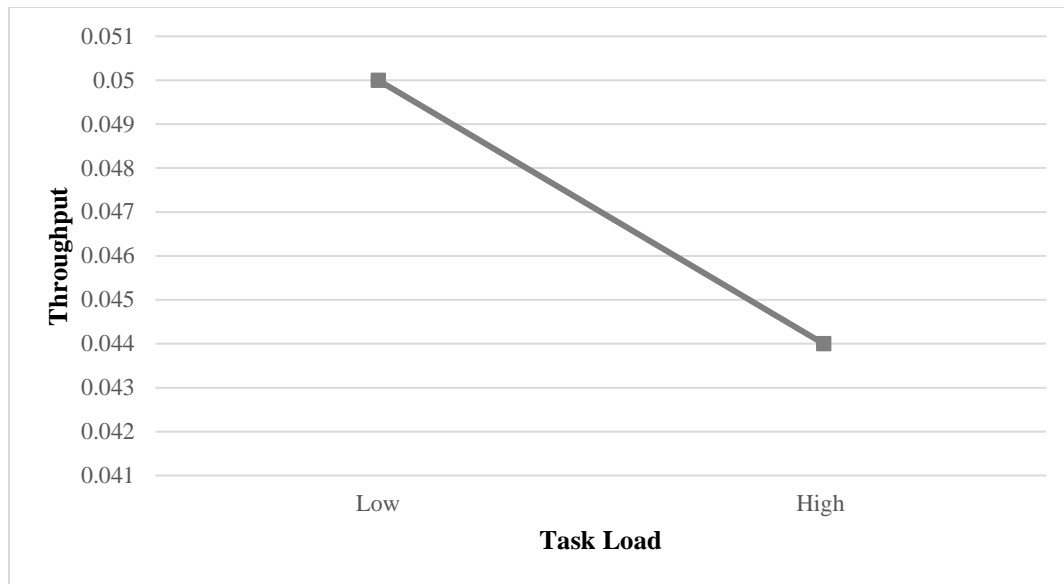


Figure 4.10. Main effect of task load on chat communications task throughput.

However, there was a statistically significant main effect of task load, $F(1, 79) = 6.897, p = 0.010$, partial $\eta^2 = 0.080$ (Figure 4.10). As participant task load increased, their performance on the chat communications task decreased. Participants' throughput for the chat communications task was lower in the high task load condition ($M = 0.044, SD = 0.002$) than the low task load condition ($M = 0.050, SD = 0.003$). The main effects of payload task complexity, $F(1, 79) = 0.096, p = 0.797$, partial $\eta^2 = 0.001$, and payload automation reliability, $F(1, 79) = 0.005, p = 0.943$, partial $\eta^2 = 0.000$, were not significant.

4.3 UAV Operator Subjective Workload

The CSS was used to assess participants' average and maximum subjective workload for each experimental block (Ames & George, 1993). Participants rated both their average and maximum workload on a seven-point scale that ranged from one (nothing to do) to seven (overloaded). For more detail on these scales, see Figure A.20 in Appendix A.

The Little's MCAR test obtained for the CSS data resulted in $\chi^2 = 28.134$ ($df = 24; p = 0.254$), which indicated the data were MCAR. Since only a small percentage of data were missing (1.2%), missing values were filled using EM.

4.3.1. Mean subjective workload. A three-way mixed ANOVA was run to understand the effects of operator task load (low and high), task environment complexity (low and high), and automation reliability (low and high) on operators’ mean subjective workload ratings. Operators rated their subjective mean workload on an anchored seven-point scale, which ranged from one (nothing to do) to seven (overloaded).

Mean workload ratings were not normally distributed, as assessed by the Kolmogorov-Smirnov test ($p < 0.05$). However, since values for skewness and kurtosis between -2 and +2 are acceptable to demonstrate normal univariate distribution (George & Mallery, 2009), the analysis was continued without data transformation for easier interpretation.

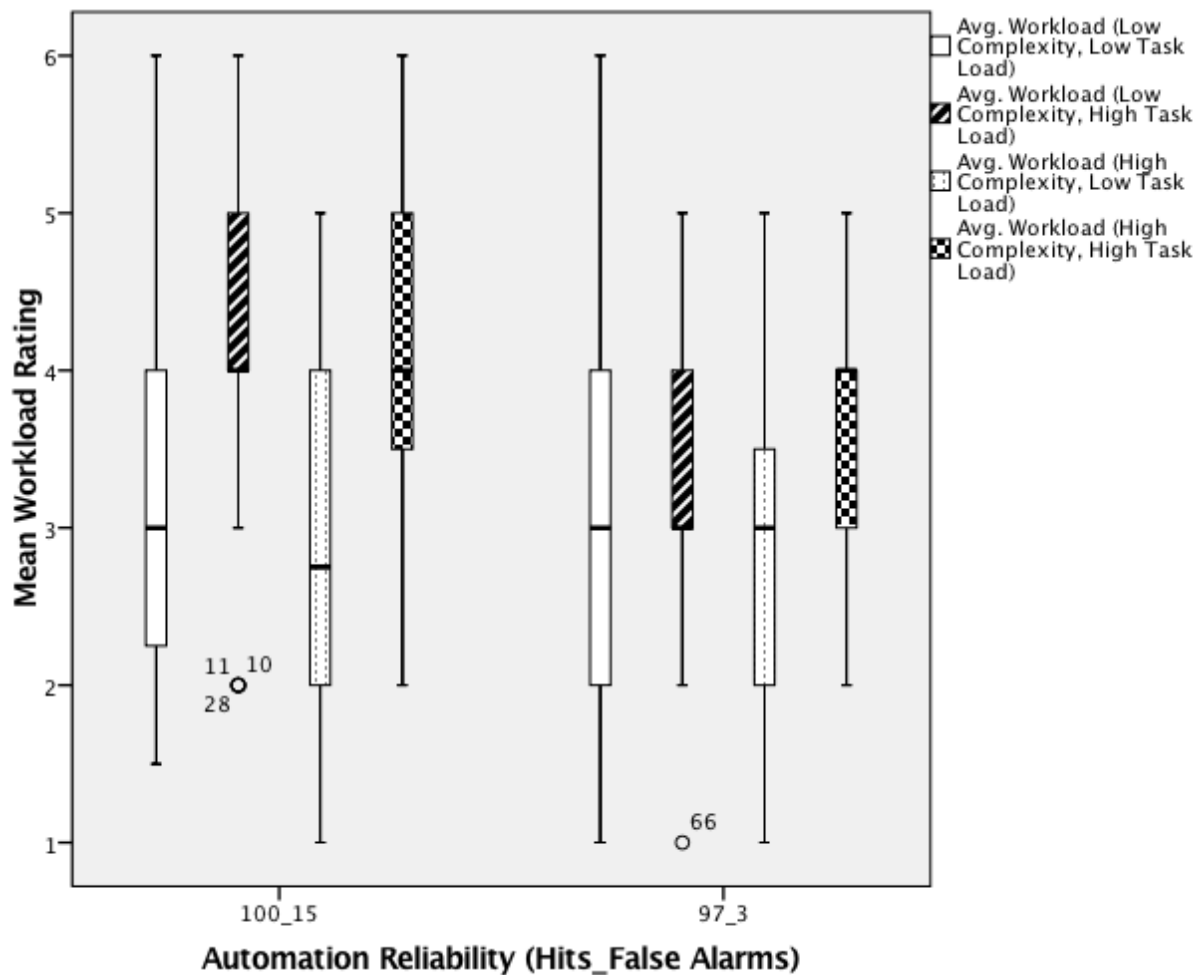


Figure 4.11. Participant workload ratings by experimental block. The workload ratings for the

unreliable automation group (100% hit rate, 15% false alarm rate) are displayed on the left and the workload ratings for reliable automation group (97% hit rate, 3% false alarm rate) are displayed on the right.

There were four outliers in the data, as assessed by inspection of a boxplot (Figure 4.11). The outliers were kept in the analysis because they were genuinely unusual values and not the result of measurement or data entry errors. The outliers did not have an appreciable effect on the analysis as assessed by a comparison of the results with and without the outliers.

The assumption of homogeneity of variances was violated for the high complexity / low task load experimental block, as assessed by Levene's test for equality of variances, $F(1,79) = 4.396, p = 0.039$. Thus, one cannot assume equal variances between groups. Mauchly's test of sphericity was not assessed because there were only two levels of both within-subjects factors. Therefore, there was only one paired difference for each and the assumption of sphericity was automatically met.

There was no statistically significant three-way interaction between reliability, task load, and complexity, $F(1, 79) = 0.000, p = 0.992, \text{partial } \eta^2 = 0.000$. There was no statistically significant simple two-way interaction between complexity and reliability, $F(1, 79) = 3.292, p = 0.073, \text{partial } \eta^2 = 0.040$. There was no statistically significant simple two-way interaction between task load and reliability, $F(1, 79) = 2.966, p = 0.089, \text{partial } \eta^2 = 0.036$. There was no statistically significant simple two-way interaction between complexity and task load, $F(1, 79) = 2.015, p = 0.160, \text{partial } \eta^2 = 0.025$.

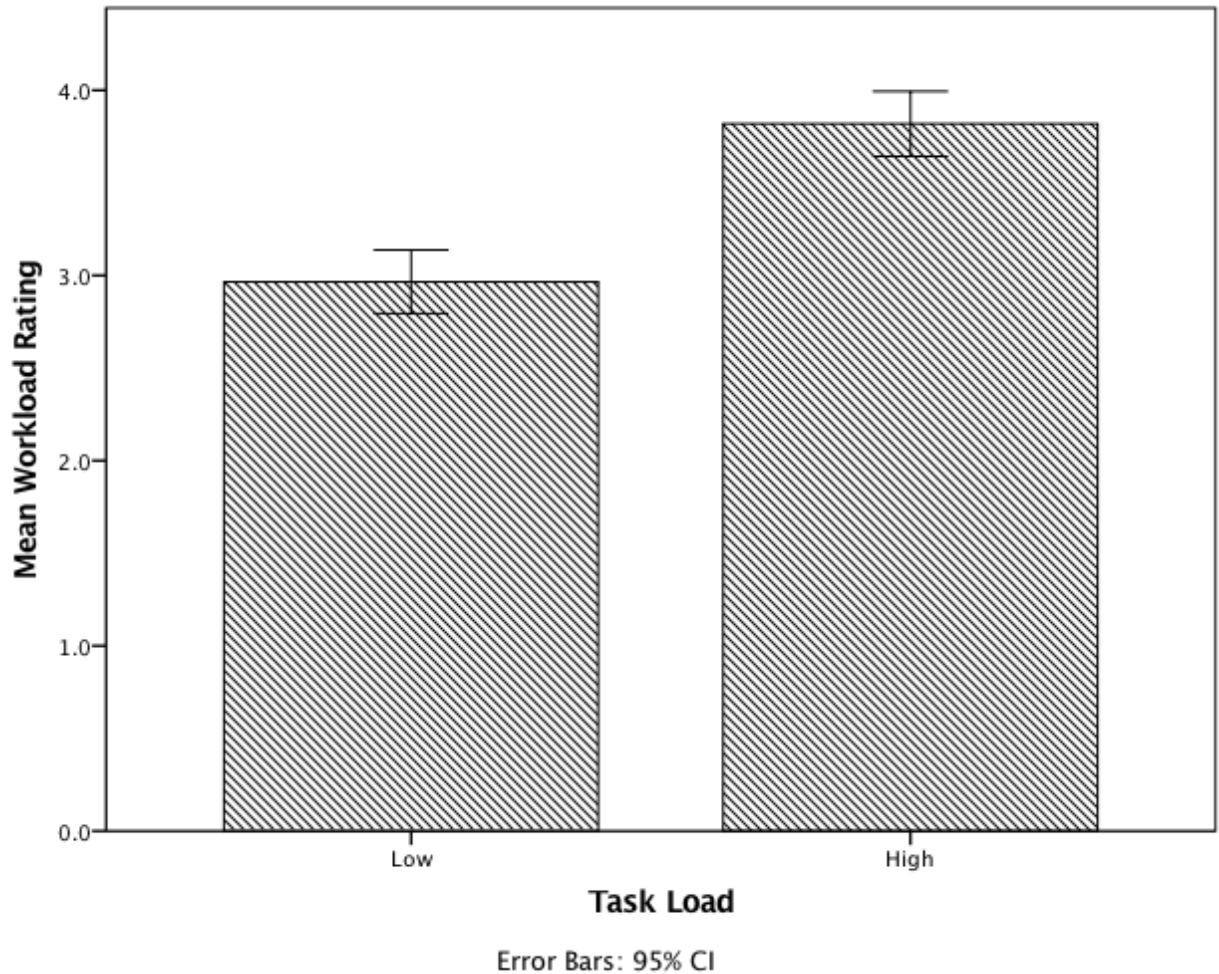


Figure 4.12. Main effect of task load on participants' mean workload ratings.

While there were no significant interactions, there was a statistically significant main effect of task load, $F(1, 79) = 57.103, p = 0.000, \text{partial } \eta^2 = 0.420$ (Figure 4.12). As participant task load increased, their average workload ratings increased. Participants' average workload ratings were significantly higher in the high task load condition ($M = 3.844, SD = 1.121$) than the low task load condition ($M = 2.966, SD = 1.010$). However, although the difference in average workload ratings between the low and high task load blocks is significant, it should be noted that the effect size is relatively small. These values correspond to a mean difference of one point on the seven-point scale and workload ratings of (4) "busy, challenging but manageable, adequate

time available” and (3) “moderate activity, easily managed, considerable spare time,” respectively.

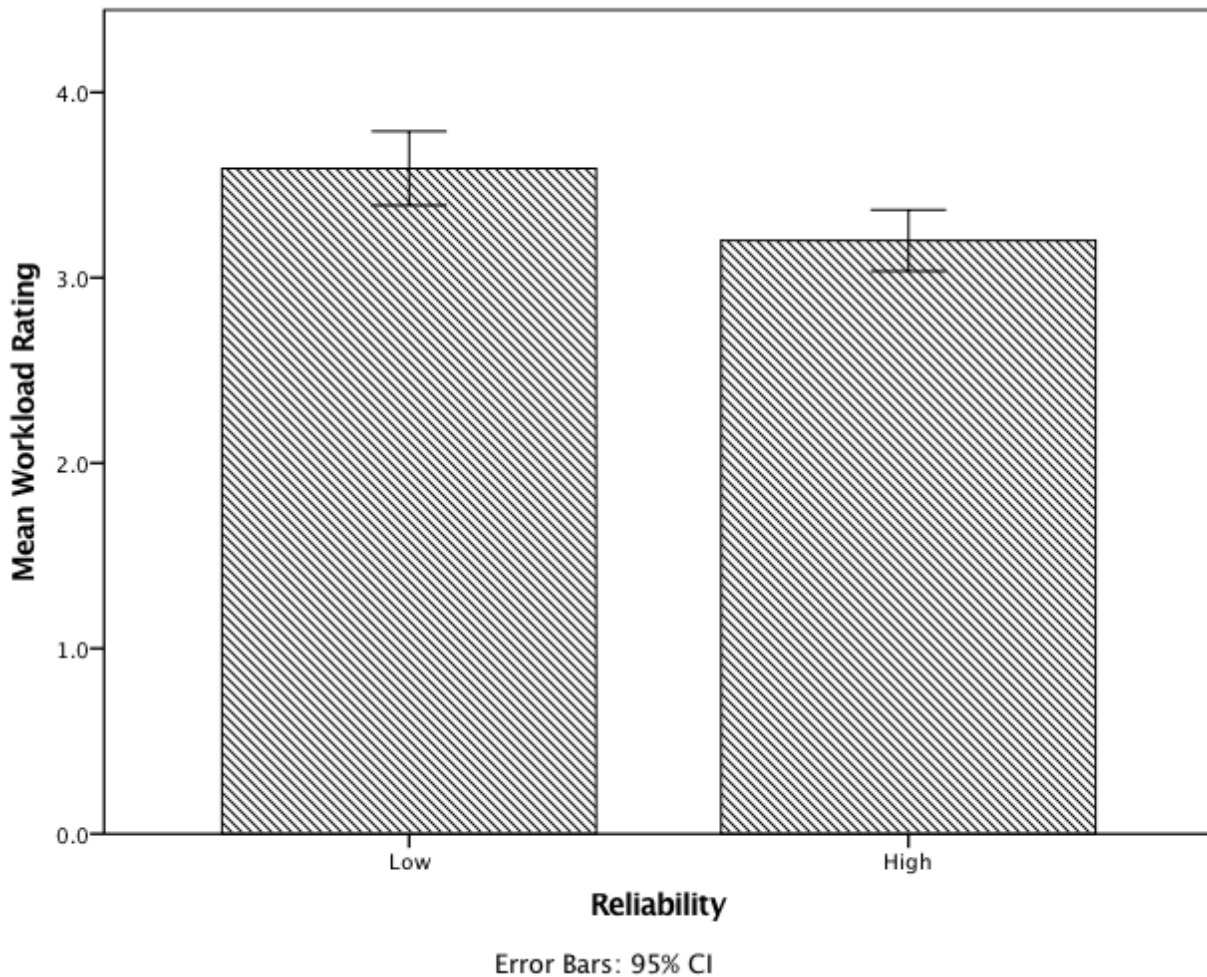


Figure 4.13. Main effect of automation reliability on participants’ mean workload ratings. The left and right bars display the results for the low (100% hit rate, 15% false alarm rate) and high (97% hit rate, 3% false alarm rate) automation reliability groups, respectively.

There was also a statistically significant main effect of reliability, $F(1, 79) = 5.609$, $p = 0.020$, partial $\eta^2 = 0.066$ (Figure 4.13). Participants’ average workload ratings were significantly higher in the low reliability automation condition ($M = 3.624$, $SD = 1.276$) than the high reliability automation condition ($M = 3.190$, $SD = 1.066$). However, as was the case for the main effect of task load, it should be noted that the effect size for the automation reliability main effect

is relatively small; the difference between the low and high reliability groups is less than half a point on the seven-point scale.

In summary, increased task load and reduced automation reliability both result in modest increases in participant's subjective mean workload.

4.3.2. Maximum subjective workload. A three-way mixed ANOVA was run to understand the effects of operator task load (low and high), task environment complexity (low and high), and automation reliability (low and high) on operators' subjective maximum workload. Operators rated their maximum workload on an anchored seven-point scale, which ranged from one (nothing to do) to seven (overloaded).

Maximum workload ratings were not normally distributed, as assessed by the Kolmogorov-Smirnov test ($p < 0.05$). However, since values for skewness and kurtosis between -2 and +2 are acceptable to demonstrate normal univariate distribution (George & Mallery, 2009), the analysis was continued without data transformation for easier interpretation.

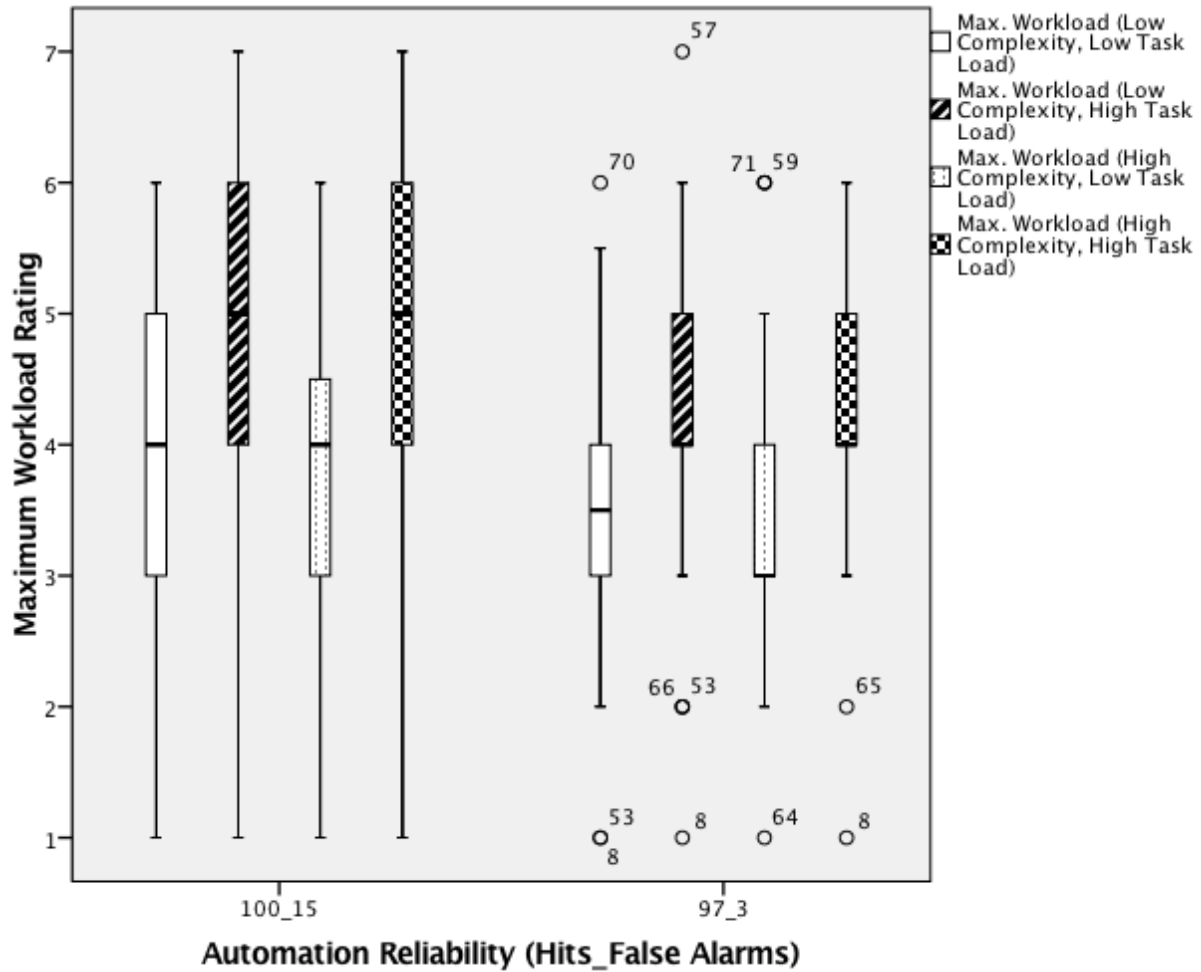


Figure 4.14. Maximum workload ratings by experimental block. The workload ratings for the unreliable automation group (100% hit rate, 15% false alarm rate) are displayed on the left and the workload ratings for reliable automation group (97% hit rate, 3% false alarm rate) are displayed on the right.

There were 12 outliers in the data, as assessed by inspection of a boxplot (Figure 4.14).

The outliers were kept in the analysis because they were genuinely unusual values and not the result of measurement or data entry errors. The outliers did not have an appreciable effect on the analysis as assessed by a comparison of the results with and without the outliers.

There was homogeneity of variances, as assessed by Levene's test for equality of variances ($p > 0.05$). Mauchly's test of sphericity was not assessed because there were only two

levels of both within-subjects factors. Therefore, there was only one paired difference for each and the assumption of sphericity was automatically met.

There was no statistically significant three-way interaction between reliability, task load, and complexity, $F(1, 79) = 1.248, p = 0.267, \text{partial } \eta^2 = 0.016$. There was no statistically significant simple two-way interaction between complexity and reliability, $F(1, 79) = 1.294, p = 0.259, \text{partial } \eta^2 = 0.016$. There was no statistically significant simple two-way interaction between task load and reliability, $F(1, 79) = 0.493, p = 0.485, \text{partial } \eta^2 = 0.006$. There was no statistically significant simple two-way interaction between complexity and task load, $F(1, 79) = 0.664, p = 0.418, \text{partial } \eta^2 = 0.008$.

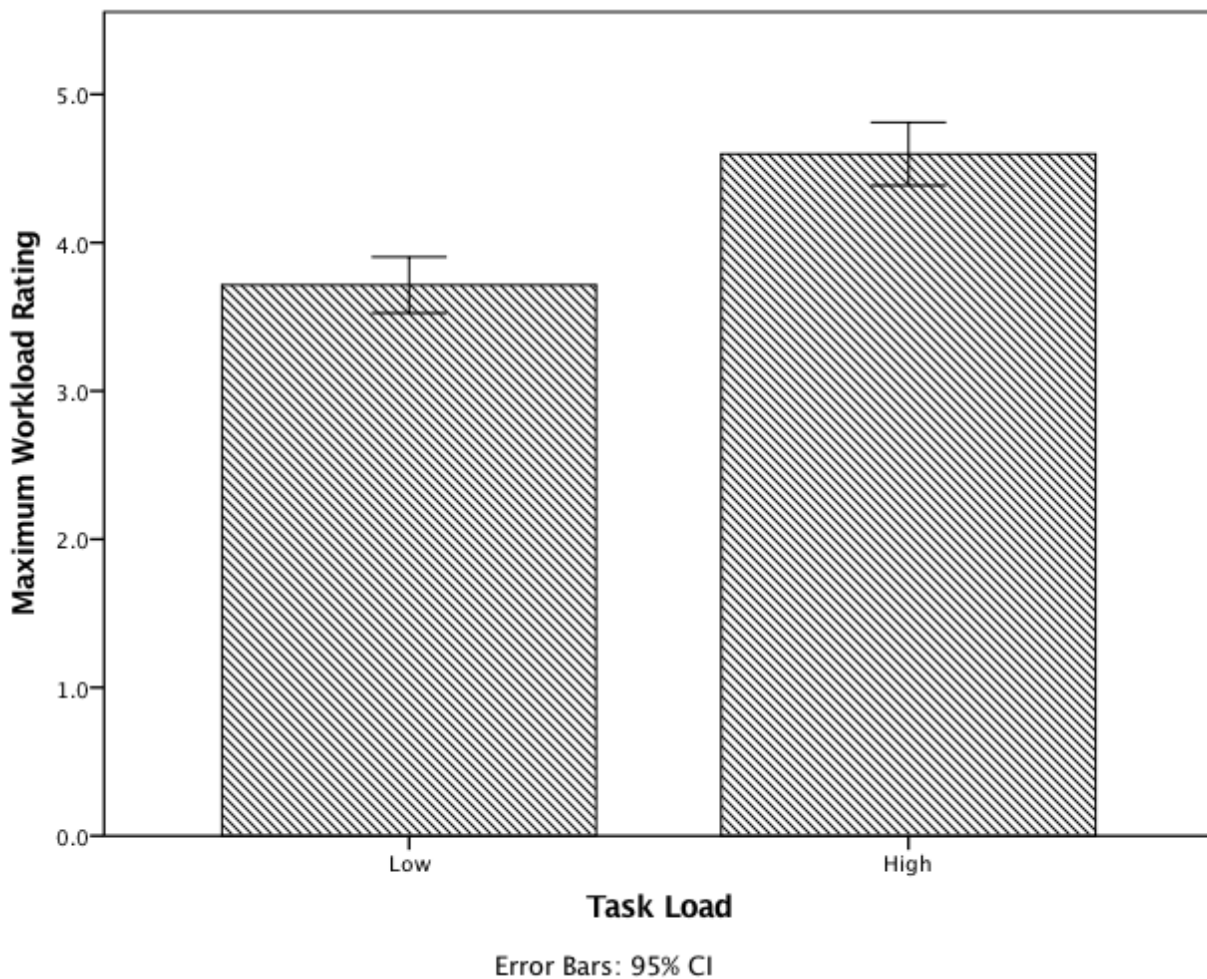


Figure 4.15. Main effect of task load on participants' maximum workload ratings.

However, there was a statistically significant main effect of task load, $F(1, 79) = 54.465$, $p = 0.000$, partial $\eta^2 = 0.408$ (Figure 4.15). As participant task load increased, their subjective maximum workload ratings increased. Participants' subjective maximum workload ratings were significantly higher in the high task load condition ($m = 4.608$, $s = 1.379$) than the low task load condition ($m = 3.702$, $s = 1.215$). These values correspond to a small, approximately one-point difference in ratings of (5) "very busy, demanding to manage, barely enough time" and (4) "busy, challenging but manageable, adequate time available," respectively.

Unlike the mean subjective workload ratings, there was no statistically significant main effect of reliability on subjective maximum workload ratings, $F(1, 79) = 3.917$, $p = 0.051$, partial $\eta^2 = 0.047$. Likewise, the main effect of complexity was not significant, $F(1, 79) = 0.022$, $p = 0.882$, partial $\eta^2 = 0.000$.

4.4 UAV Operator Subjective Fatigue

In addition to workload, the CSS was also used to assess participants' subjective fatigue for each experimental block (Ames & George, 1993). Participants rated their fatigue on a seven-point scale that ranged from one (fully alert) to seven (completely exhausted). Please see Figure A.19 in Appendix A for more detail on the fatigue rating scale. As stated above, the small percentage of missing CSS data (1.2%) was imputed using EM.

A three-way mixed ANOVA was run to understand the effects of operator task load (low and high), task environment complexity (low and high), and automation reliability (low and high) on operators' subjective fatigue.

Fatigue ratings were not normally distributed, as assessed by the Kolmogorov-Smirnov test ($p < 0.05$). However, since values for skewness and kurtosis between -2 and +2 are

acceptable to demonstrate normal univariate distribution (George & Mallery, 2009), the analysis was continued without data transformation for easier interpretation.

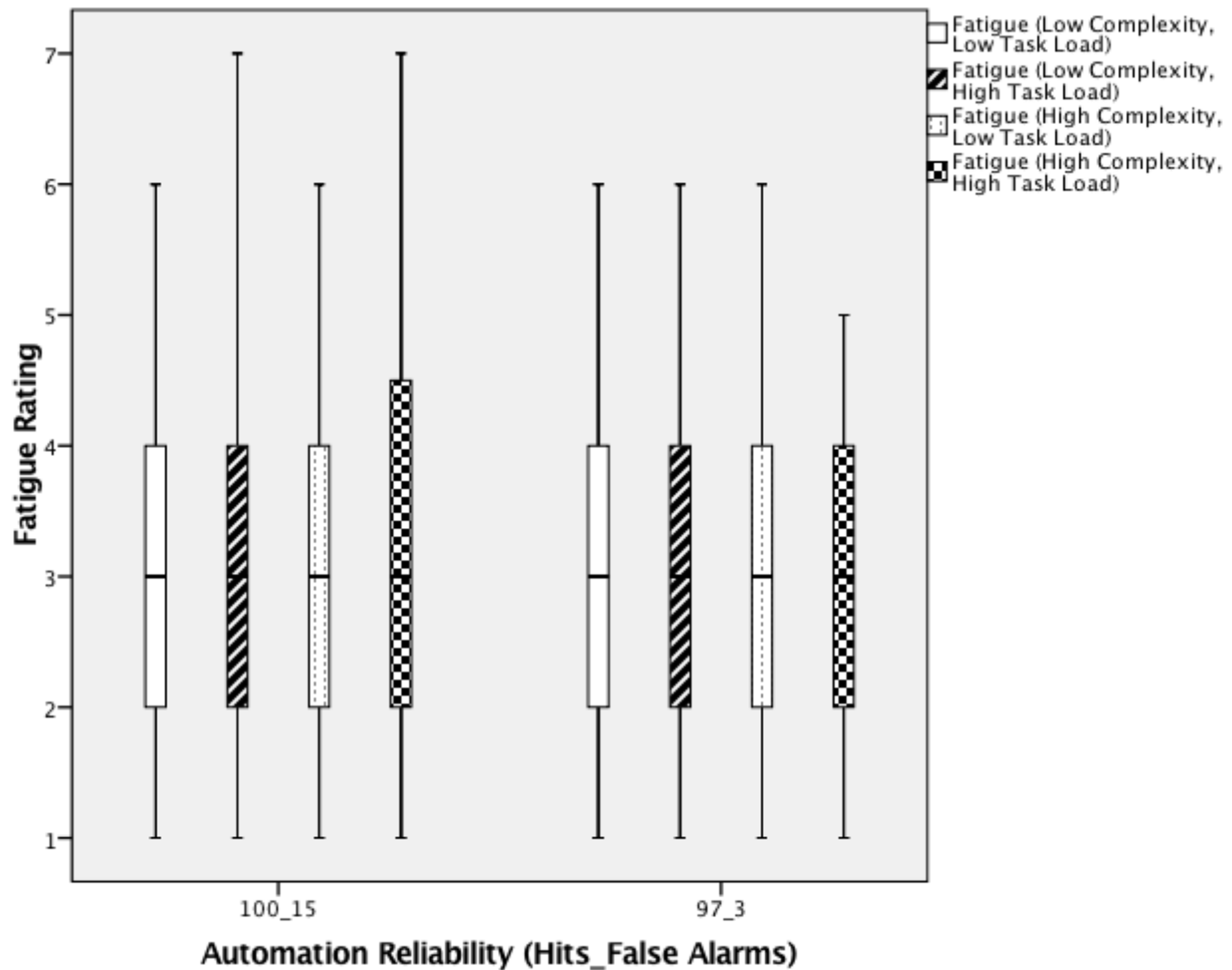


Figure 4.16. Fatigue ratings by experimental block. The fatigue ratings for the unreliable automation group (100% hit rate, 15% false alarm rate) are displayed on the left and the fatigue ratings for reliable automation group (97% hit rate, 3% false alarm rate) are displayed on the right.

Inspection of a boxplot indicated that there were no outliers in the data (Figure 4.16). The assumption of homogeneity of variances was violated for the high complexity / high task load experimental block, as assessed by Levene's test for equality of variances, $F(1,79) = 4.931$, $p = 0.029$. Thus, one cannot assume equal variances between groups. Mauchly's test of sphericity was not assessed because there are only two levels of both within-subjects factors. Therefore,

there is only one paired difference for each and the assumption of sphericity is automatically met.

There was no statistically significant three-way interaction between reliability, task load, and complexity, $F(1, 79) = 0.023, p = 0.880, \text{partial } \eta^2 = 0.000$. There was no statistically significant simple two-way interaction between complexity and reliability, $F(1, 79) = 0.597, p = 0.442, \text{partial } \eta^2 = 0.008$. There was no statistically significant simple two-way interaction between task load and reliability, $F(1, 79) = 1.202, p = 0.276, \text{partial } \eta^2 = 0.015$. There was no statistically significant simple two-way interaction between complexity and task load, $F(1, 79) = 0.757, p = 0.387, \text{partial } \eta^2 = 0.009$.

Unlike the workload ratings, there was no statistically significant main effect of task load on fatigue ratings, $F(1, 79) = 1.817, p = 0.182, \text{partial } \eta^2 = 0.022$. The main effects for reliability, $F(1, 79) = 0.014, p = 0.840, \text{partial } \eta^2 = 0.001$, and complexity, $F(1, 79) = 1.324, p = 0.253, \text{partial } \eta^2 = 0.016$, were also not significant. Task load, payload task complexity, and the reliability of the payload task automation did not affect participants' experience of fatigue.

4.5 UAV Operator Automation Dependence

Participants' degree of automation dependence was defined as the percentage of their responses that followed the recommendation of the payload automation. A lower percent agreement score indicated that a participant more frequently overrode the potential targets selected by the automation or manually selected additional potential targets.

The Little's MCAR test obtained for these variables resulted in $\chi^2 = 26.012 (df = 17; p = 0.074)$, which indicated that the data were MCAR. In other words, there was no pattern to the missing data. However, listwise deletion of missing values was not ideal since 10.2% of the data were missing (Little, 1988). This missing data can be attributed to blocks where, due to planning

decisions made in that block and the prior block, no searches took place. None of the missing data was due to attrition.

Participants' percent agreement was not normally distributed, as assessed by the Kolmogorov-Smirnov test ($p < 0.05$). A log transformation was applied to the data to correct its negative skew (Tabachnick and Fidell, 2013). After transformation, the skewness and kurtosis values fell between the -2 and +2 range acceptable to demonstrate normal univariate distribution (George & Mallery, 2009).

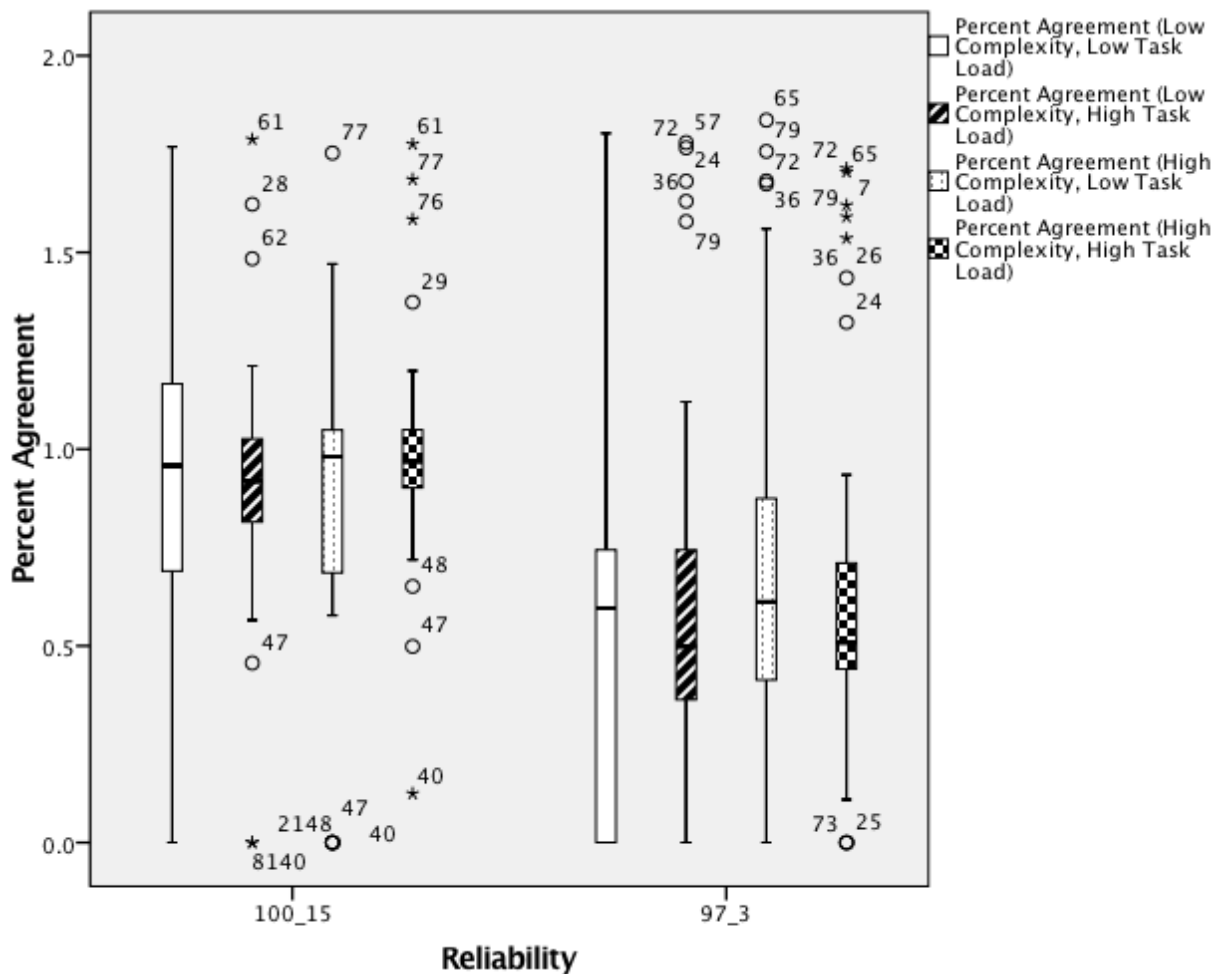


Figure 4.17. Participant percent agreement with automation by experimental block. The percent agreement of participants in the unreliable automation group (100% hit rate, 15% false alarm rate) is displayed on the left and the percent agreement of participants in the reliable automation group (97% hit rate, 3% false alarm rate) is displayed on the right. This data was log transformed.

There were 36 outliers in the data, as assessed by inspection of a boxplot (Figure 4.17). The outliers were kept in the analysis because they were genuinely unusual values and not the result of measurement or data entry errors. The outliers did not have an appreciable effect on the analysis as assessed by a comparison of the results with and without the outliers.

An REML LMM analysis was run to assess the log transformed percent agreement data as a function of task load, task complexity, and automation reliability. As with prior analyses, the LMM approach was used instead of a mixed ANOVA because LMM is able to accommodate missing data.

$$\text{Percent Agreement} \sim \text{Task Load} + \text{Task Complexity} + \text{Automation Reliability} + (1|\text{Participant}) + (1|\text{Block}) + \varepsilon \quad (4.3)$$

Task load, task complexity, and automation reliability (and their two-way interaction terms) were entered into the model as fixed factors. Participant ID and experimental block were included in the model as random factors (Magezi, 2015; Winter, 2013). This information is summarized in Equation 4.3, which also includes the general error term “ ε .”

Levene’s test was significant for the high complexity, low task load experimental block ($p > 0.05$), so homogeneity of variances cannot be assumed.

Multiple models (compound symmetry; first order autoregressive, AR(1); and unstructured) were fit via a penalized likelihood approach or, more specifically, through comparison of their BIC and AIC values (Seltman, 2018). The selected model utilized an unstructured covariance structure since its BIC and AIC values were smaller (Seltman, 2018). Mauchly’s test of sphericity was not assessed because the model did not assume sphericity or compound symmetry. However, even if sphericity was assumed, there were only two levels of

both within-subjects factors. Therefore, there would only be one paired difference for each and the assumption of sphericity would automatically be met.

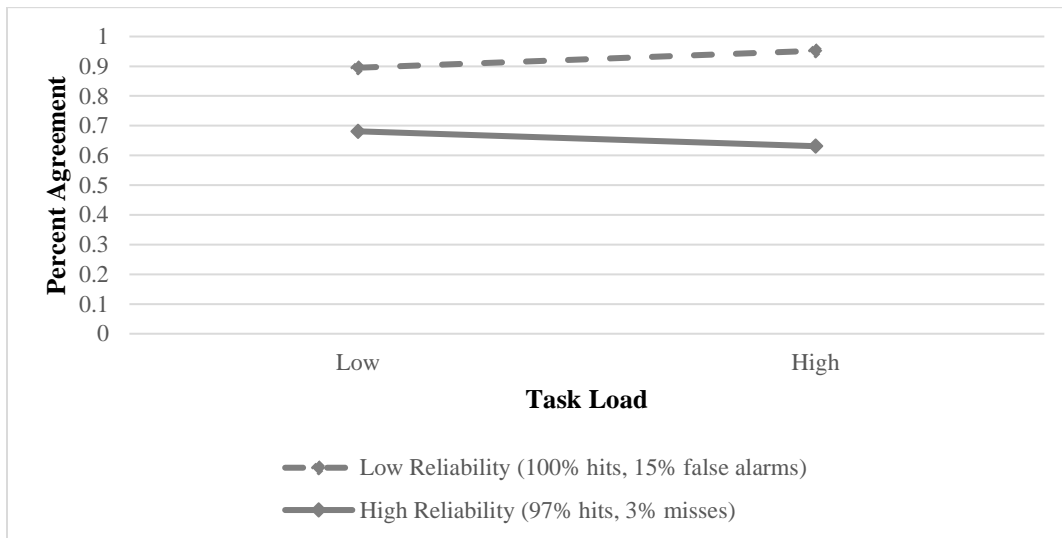


Figure 4.18. Reliability and task load interaction for participants' percent agreement with the payload task automation. This data was log transformed.

There was a statistically significant simple two-way interaction between reliability and task load, $F(1, 69.242) = 4.035, p = 0.048, \text{partial } \eta^2 = 0.055$ (Figure 4.18). There were no statistically significant simple two-way interactions between reliability and complexity, $F(1, 77.530) = 0.090, p = 0.764, \text{partial } \eta^2 = 0.001$, or complexity and task load, $F(1, 68.164) = 1.153, p = 0.287, \text{partial } \eta^2 = 0.017$.

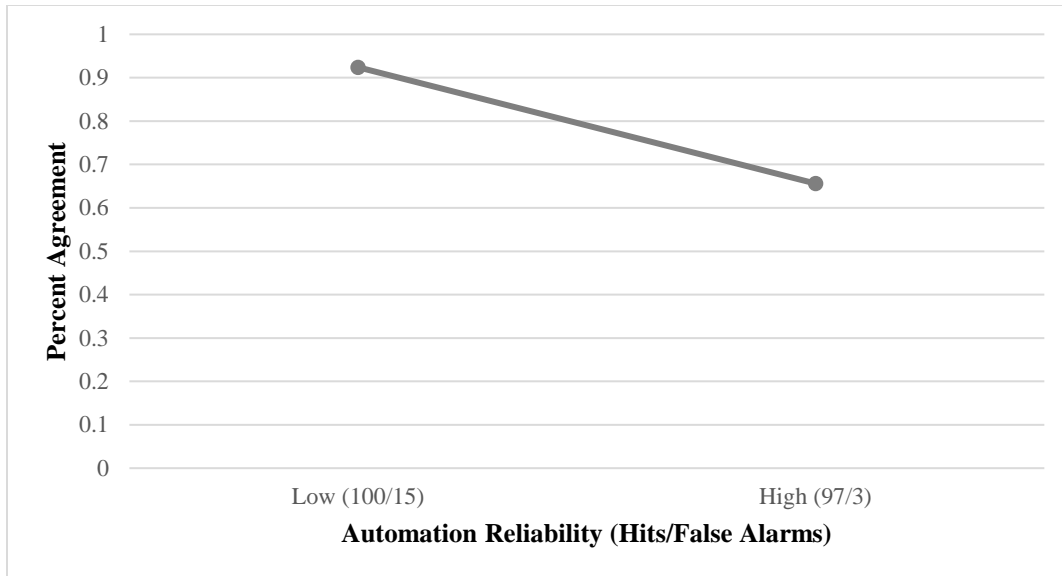


Figure 4.19. Effect of automation reliability on participants' percent agreement with the automation. This data was log transformed.

There was also a significant main effect of automation reliability, $F(1, 76.563) = 10.128$, $p = 0.002$, partial $\eta^2 = 0.117$ (Figure 4.19). However, neither the main effect for task load, $F(1, 68.968) = 0.020$, $p = 0.887$, partial $\eta^2 = 0.000$, nor the main effect for payload task complexity, $F(1, 71.539) = 0.535$, $p = 0.467$, partial $\eta^2 = 0.007$, were significant.

Overall, participants agreed with the unreliable automation more often ($M = 0.917$, $SD = 0.379$) than the reliable automation ($M = 0.640$, $SD = 0.488$), a surprising finding given that participants were penalized for false alarms and the unreliable automation had a liberal response criterion (i.e., it was prone to false alarms). However, there was also a significant interaction of automation reliability and task load that should be considered.

Although participants generally tended to rely on the unreliable, false alarm-prone automation more often, task load had more bearing on their reliance on said unreliable automation relative to the reliable automation. While the effect size of this interaction was quite small, participant reliance on the reliable automation was slightly more stable between periods of low ($M = 0.681$, $SD = 0.067$) and high ($M = 0.631$, $SD = 0.057$) task load. Conversely,

participants working in conjunction with the unreliable automation relied on it slightly more often when task saturated ($M = 0.952$, $SD = 0.057$), but tended to override it slightly less often when their task load was low ($M = 0.895$, $SD = 0.068$).

4.6 UAV Operator Trust and Self-Confidence in Automation

A brief four-question Likert-type survey was administered at the end of the study. Responses were missing for one participant ($n = 80$). The survey, modeled after Lee and Moray's validated (1994) scale, asked participants to rate their trust in the payload task automation and their self-confidence that they could manually perform the same target search task. The purpose of this survey was to gauge whether participants in the low and high automation reliability groups noticed a difference in the reliability of the payload task automation and if their trust and self-confidence were accordingly impacted.

4.6.1. Trust rating. Participants were asked: "To what extent did you trust (i.e., believe in the accuracy of) the sensor feed automation to select correct targets and enclose them in brown boxes in this scenario?" An independent samples t-test was run to assess whether there was a significant difference in participants' trust of the sensor feed automation in the low and high automation reliability conditions.

There were no outliers in the data, as assessed by inspection of a boxplot. The data was not normally distributed, as assessed by the Kolmogorov-Smirnov test ($p < 0.05$). However, since values for skewness and kurtosis between -2 and +2 are acceptable to demonstrate normal univariate distribution (George & Mallery, 2009), and the independent-samples t-test is fairly robust to deviations from normality, the analysis was continued without data transformation for easier interpretation. There was homogeneity of variances, as assessed by Levene's test for equality of variances ($p = .938$).

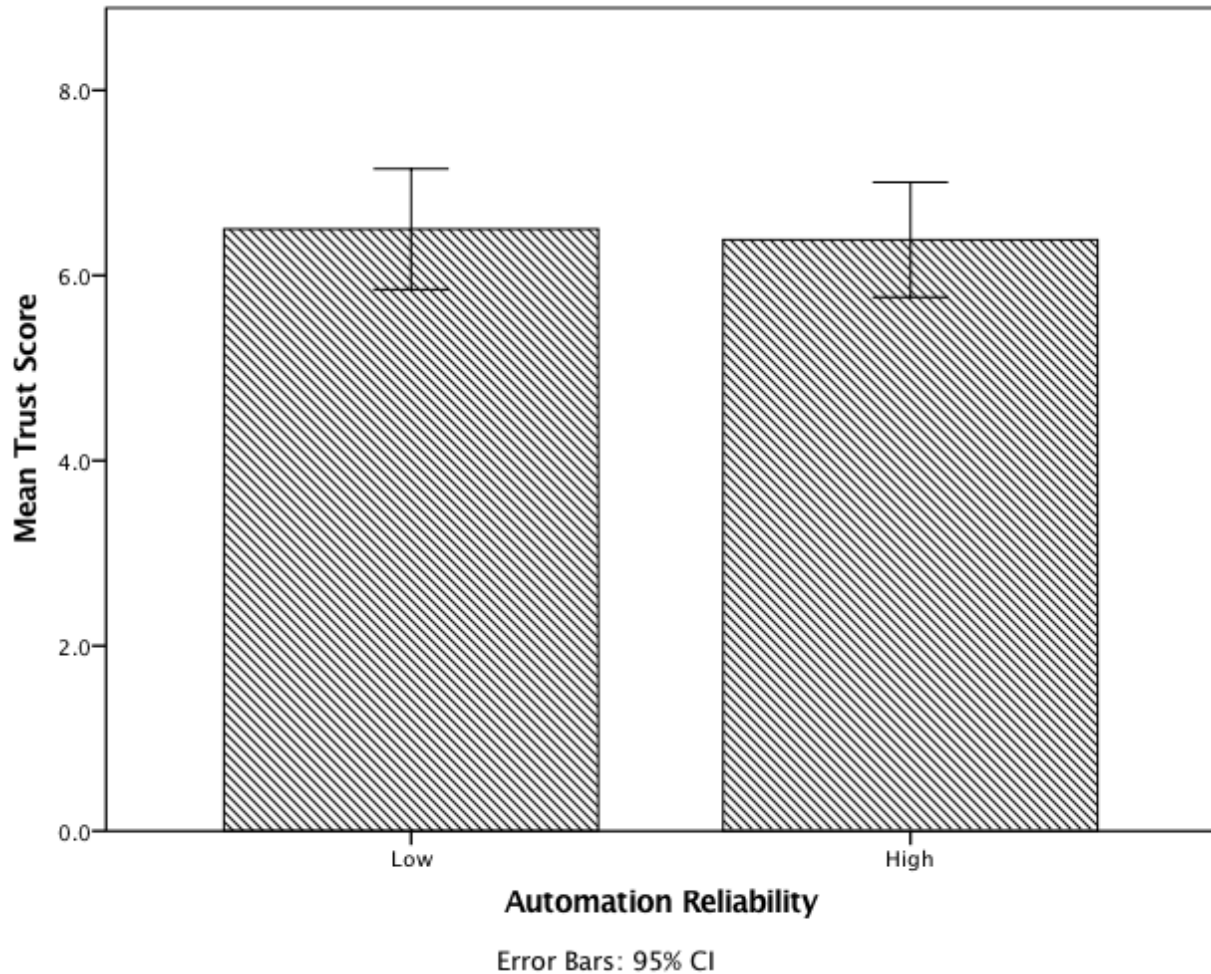


Figure 4.20. Mean trust ratings of participants by automation reliability. The left and right bars display the results for the low (100% hit rate, 15% false alarm rate) and high (97% hit rate, 3% false alarm rate) automation reliability groups, respectively.

Although participants in the low reliability condition ($M = 6.500$, $SD = 2.013$) trusted the automation to select correct targets and enclose them in brown boxes slightly more than participants in the high reliability condition ($M = 6.385$, $SD = 1.969$), the difference between the two groups was not statistically significant, $M = 0.115$, 95% CI $[-0.772, 1.00]$, $t(78) = 0.257$, $p = .797$, $d = 0.058$ (Figure 4.20).

4.6.2. Reliance rating. Participants were asked: “To what extent did you rely (i.e., actually use) the automatically selected targets in this scenario?” An independent samples t-test

was run to assess whether there was a significant difference in participants' perceived reliance on the sensor feed automation in the low and high automation reliability conditions.

There were three outliers in the data, as assessed by inspection of a boxplot. The outliers were kept in the analysis because they were genuinely unusual values and did not have an appreciable effect on the analysis. The data was not normally distributed, as assessed by the Kolmogorov-Smirnov test ($p < 0.05$). However, since values for skewness and kurtosis between -2 and +2 are acceptable to demonstrate normal univariate distribution (George & Mallery, 2009), and the independent-samples t-test is fairly robust to deviations from normality, the analysis was continued without data transformation for ease of interpretation. There was homogeneity of variances, as assessed by Levene's test for equality of variances ($p = 0.631$).

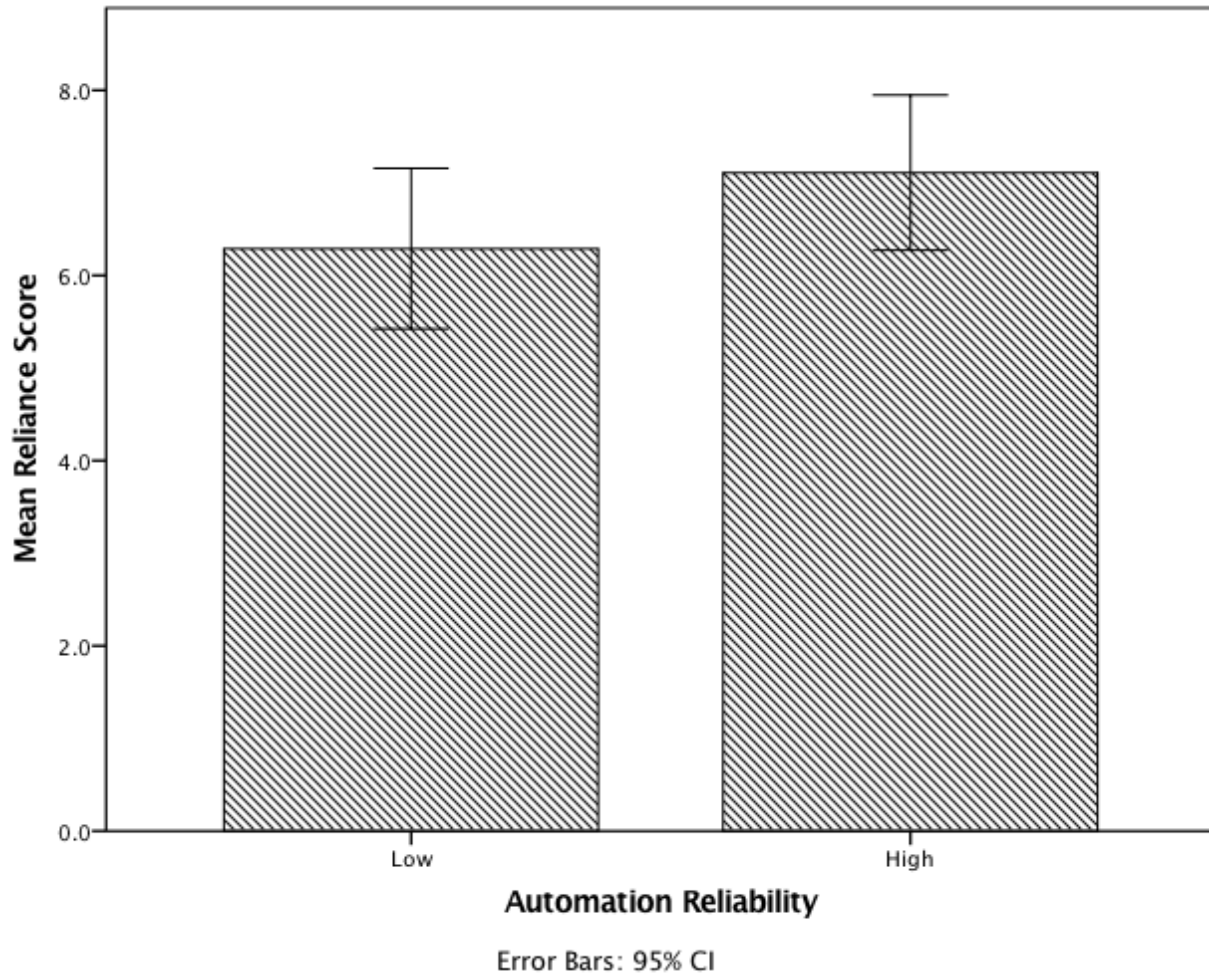


Figure 4.21. Mean reliance ratings of participants by automation reliability. The left and right bars display the results for the low (100% hit rate, 15% false alarm rate) and high (97% hit rate, 3% false alarm rate) automation reliability groups, respectively.

Participants in the high reliability condition ($M = 7.110$, $SD = 2.652$) reported relying on the automation more to select correct targets than those in the low reliability condition ($M = 6.290$, $SD = 2.672$). However, the difference between the two groups was not statistically significant, $M = -0.820$, 95% CI $[-2.005, 0.365]$, $t(78) = -1.378$, $p = .631$, $d = 0.308$ (Figure 4.21).

4.6.3. Self-confidence rating. Participants were asked: “To what extent were you self-confident that you could successfully select all the correct targets that appear in your sensor feeds if they were not pre-selected and enclosed in brown boxes by the automation in this scenario?”

An independent samples t-test was run to assess whether there was a significant difference in participants' self-confidence that they could manually perform the automated task (i.e., select targets) between the low and high automation reliability conditions.

Inspection of a boxplot revealed no outliers in the data and the data were normally distributed, as assessed by the Kolmogorov-Smirnov test ($p > 0.05$). There was homogeneity of variances, as assessed by Levene's test for equality of variances ($p = .859$).

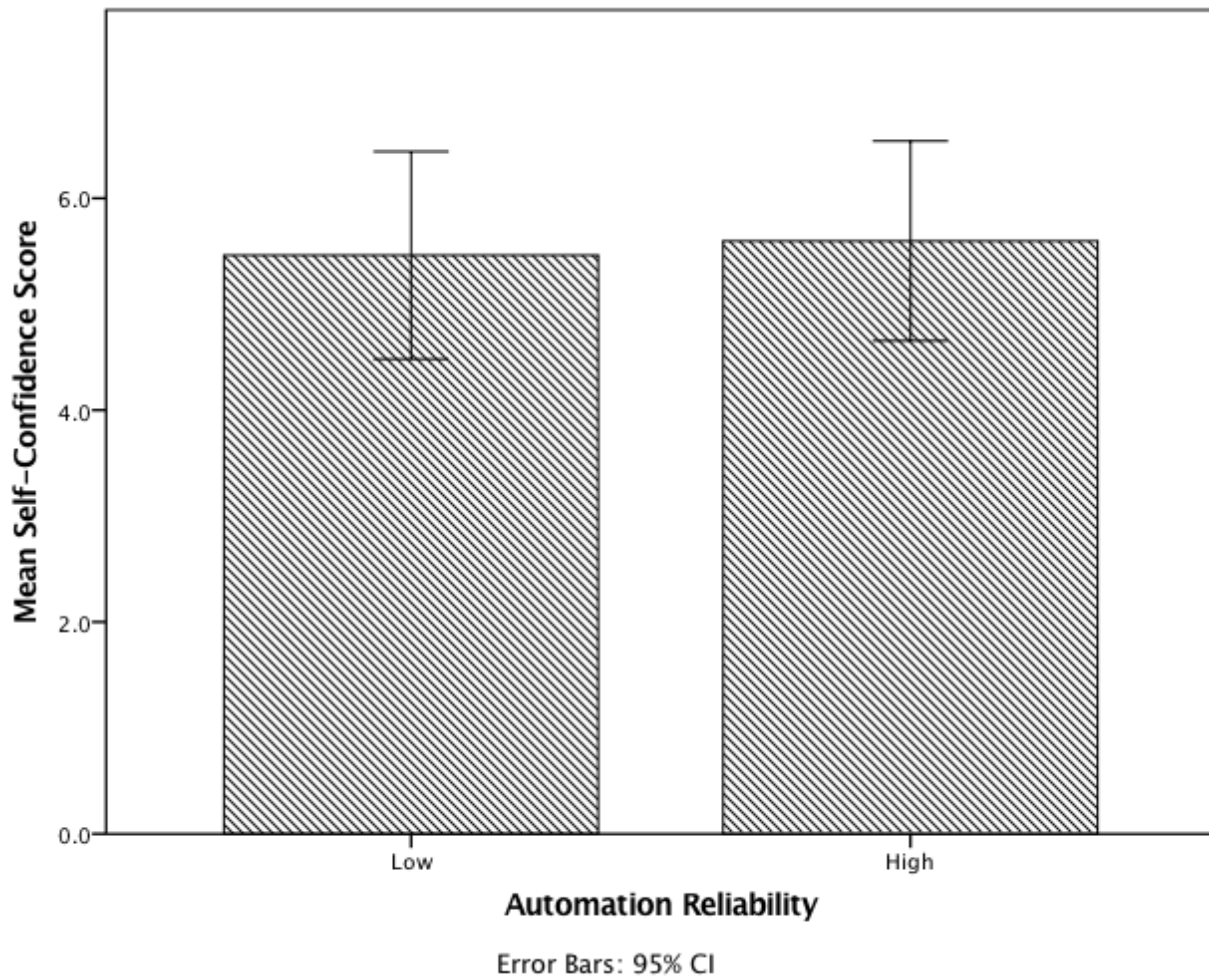


Figure 4.22. Mean self-confidence ratings of participants by automation reliability. The left and right bars display the results for the low (100% hit rate, 15% false alarm rate) and high (97% hit rate, 3% false alarm rate) automation reliability groups, respectively.

Participants in the high reliability condition ($M = 5.598$, $SD = 2.984$) were slightly more self-confident that they could select targets manually than participants in the low reliability

condition ($M = 5.462$, $SD = 3.023$). However, the difference between the two groups was not statistically significant, $M = -0.136$, 95% CI [-1.473, 1.201], $t(78) = -0.203$, $p = .859$, $d = 0.045$ (Figure 4.22).

4.6.4. Perceived performance improvement rating. Participants were asked: “To what extent do you think the sensor feed automation pre-selecting targets and enclosing them in brown boxes improved your performance in this scenario compared to your performance if you were to select all targets manually?” A Welch t-test was run to assess whether there was a significant difference in participants’ perceived performance improvement due to the sensor feed automation (relative to manual performance) between the low and high automation reliability conditions.

There was one outlier in the data, as assessed by inspection of a boxplot. The outlier was kept in the analysis because it was a genuinely unusual value and did not have an appreciable effect on the analysis. The data was not normally distributed, as assessed by the Kolmogorov-Smirnov test ($p < 0.05$). However, since values for skewness and kurtosis between -2 and +2 are acceptable to demonstrate normal univariate distribution (George & Mallery, 2009), and the independent-samples t-test is fairly robust to deviations from normality, the analysis was continued without data transformation for ease of interpretation. A Welch t-test was employed because the assumption of homogeneity of variances was violated, as assessed by Levene's test for equality of variances, $p = .006$ (Howell, 2010; Welch, 1947).

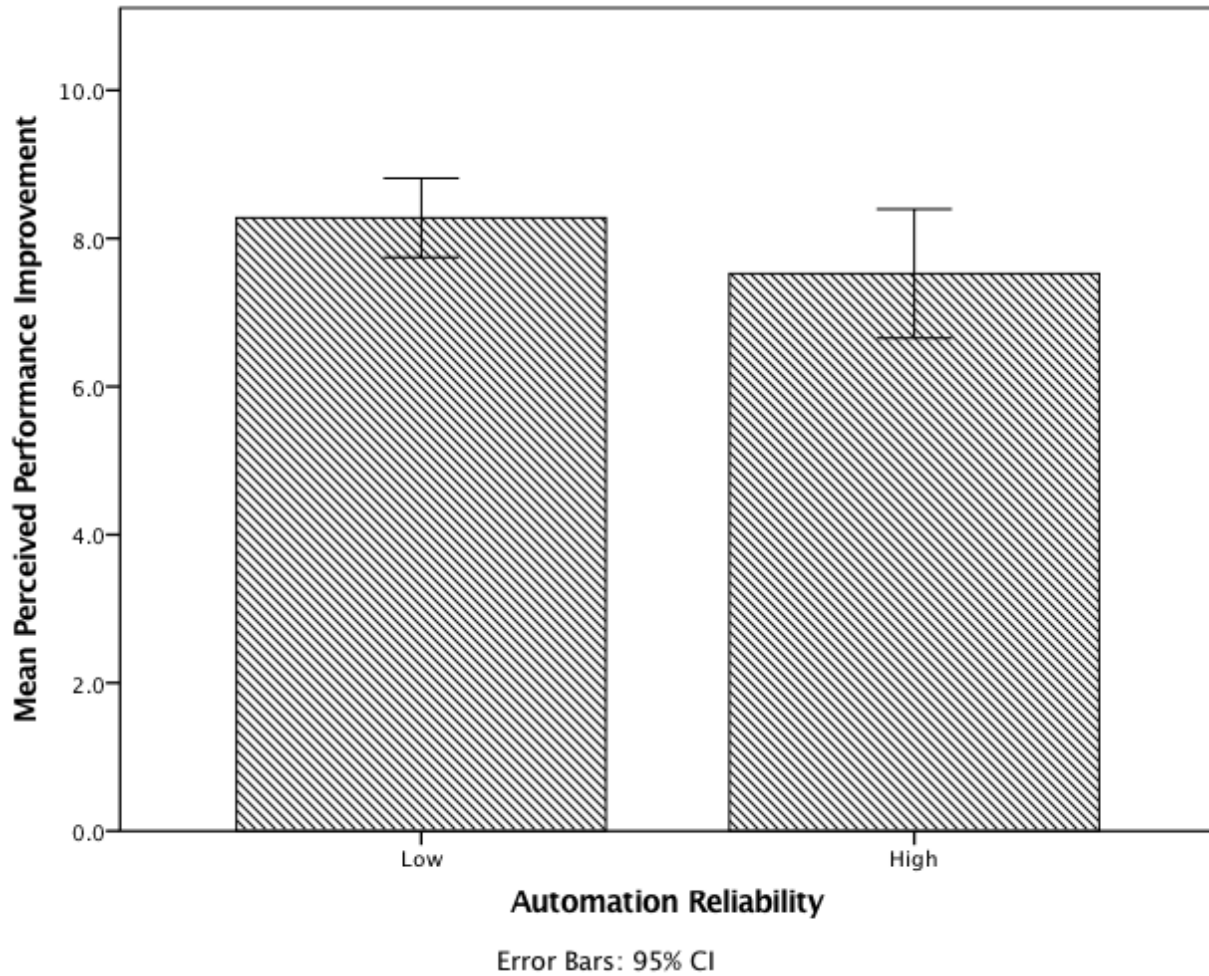


Figure 4.23. Mean perceived improvement ratings of participants by automation reliability. The left and right bars display the results for the low (100% hit rate, 15% false alarm rate) and high (97% hit rate, 3% false alarm rate) automation reliability groups, respectively.

Participants in the low reliability condition ($M = 8.276$, $SD = 1.626$) felt that the payload task automation improved their performance relative to fully manual performance more than participants in the high reliability condition ($M = 7.524$, $SD = 2.754$). However, the difference between the two groups was not statistically significant, $M = -0.752$, 95% CI $[-.2557, 1.760]$, $t(65.70) = 1.490$, $p = .141$, $d = 0.333$ (Figure 4.23).

5 Discussion

5.1 Overview

Numerous studies have examined the effects of task load and automation reliability on operators' UAV supervisory control performance, subjective workload, and automation dependence. Although comparatively few, there have also been studies investigating the effects of task complexity on UAV supervisory control performance and automation dependence. However, the present study is the first to examine the combined effects of these three factors—task load, task environment complexity, and automation reliability—on operator performance, subjective workload, and automation dependence. Furthermore, the present study examined the effects of these factors on a very unique population: student naval aviators (SNAs) and Student Naval Flight Officers (SNFOs). SNAs/SNFOs tend to be a highly homogenous and range-restricted population relative to undergraduate university students, who are the usual subjects of such research.

5.2 Effects of Task Load

The first goal of the present study was to determine whether differences in task load, task environment complexity, and automation reliability affected participants' performance. Participants' UAV supervisory control performance was assessed using three performance metrics, one for each of the three primary subtasks within the SCOUT test bed. First, participants' UAV routing performance was assessed by their adjusted expected value on the task. Second, participants' target identification performance on the automated payload task was assessed by their percent accuracy. Finally, participants' performance on the chat communications task was assessed by the throughput of their responses to messages from Command and Intelligence.

Overall, participant task load significantly impacted their performance on all three of the subtasks. High task load resulted in overall reduced performance on the automated payload and chat communications tasks. In contrast, high task load resulted in improved UAV routing performance. The improved UAV routing performance could be due to the relatively low frequency of new targets, even in the high task load condition, relative to chat messages and payload task automation errors. In each high task load block, three new targets appeared versus 32 chat messages. Thus, even in the high task load condition, the UAV routing task was relatively manageable. In the low task load condition, only one new target appeared, which could have resulted in reduced performance consistent with the parabolic utilization-performance curve analogous to the Yerkes-Dodson relationship (Cummings & Nehme, 2009).

These results could also reflect a multiple resource theory (MRT) model of workload (Wickens, 1984, 2002, 2008), which would assume that the UAV routing task employed different information processing resources than the concurrent chat communications and payload tasks, thereby facilitating efficient parallel processing and task sharing. However, since all of the SCOUT subtasks relied heavily on visual modalities of processing, the results are more easily explained by single resource theory (SRT) and reduced resource demand: participants offloaded some of their tasking onto the payload automation to increase their available resources, and therefore performance, on the UAV supervisory control task and this effect was magnified in the high task load condition. Dixon, Wickens, and Chang (2003) found that offloading tasking onto automation improved UAV routing performance in a dual-task environment.

In addition, even though all subtasks were presented with equal importance during training, it is possible that participants allocated more resources (i.e., effort) to the UAV routing task during the high task load blocks because, in some ways, it was the “unofficial primary task.”

First, performance on the UAV routing task directly impacted performance on the concurrent payload task. A participant could not search a target if they did not first assign a UAV to that target, and poor route planning resulted in fewer active searches and thus fewer opportunities to earn points. Second, the UAV task was the most salient task: it took up the majority of the screen, it was colorful, it involved the presentation of attention-grabbing stimuli (e.g., surfaced enemy submarines), and it included the most direct representation of the UAVs themselves. Research has demonstrated a moderate to strong positive linear relationship between task salience and perceived task importance (Colvin, Funk, & Braune, 2005). Relatively stable primary task performance and reduced “secondary” (chat communications and payload task) performance resulting from increased task load is consistent with the literature (Dickson, Wickens, & Chang, 2003). Finally, the UAV routing task was anecdotally considered more engaging by participants due to its “puzzle-like” nature, so there could have been a “fun factor” that contributed to participant prioritization of the UAV routing task.

All that being said, interpretation of the main effect of task load on participant performance might be misleading, as the multiple significant interactions between task load, task complexity, and automation reliability suggest. While payload task complexity and the reliability of the payload task automation did not seem to affect participants’ UAV-routing performance, the same could not be said for the payload task and the chat communications task.

Both increased task load and decreased automation reliability resulted in a significant decrease in participant performance on the automated payload task. In addition, there was a significant interaction between payload task reliability and task load on participants’ payload task performance, with unreliable automation associated with a more marked performance decrement

when participant task load was high. The availability of reliable automation seemed to partially mitigate the impact of increased task load on participant payload task performance.

5.3 Effects of Automation Reliability

While increased automation reliability improved participant performance on the automated subtask itself, it also affected concurrent task performance, namely performance on the chat communications task. There was a significant three-way interaction between participants' overall task load, the complexity of the payload task, and the reliability of the payload task automation on participants' chat communication performance. When payload task automation was unreliable (due to a liberal response criterion), participants' communication performance was largely unaffected by changes in task load as long as the concurrent payload task remained relatively simple (although participants communicated slightly better when overall task load was high). When the payload task became more complex and the payload task automation remained unreliable, however, there was a more pronounced performance drop on the communication task when overall task load increased. On the other hand, when the payload task was relatively simple and payload task automation was reliable, there was a more pronounced performance drop on the communication task when overall task load increased. When the payload task became more complex and the payload task automation remained reliable, however, participants' communication performance was largely unaffected by changes in task load (although participants communicated slightly better when overall task load was high).

Overall, when the payload task automation was less reliable, there was a pronounced decrease in performance on the concurrent communication task attributable to increased overall task load when the complexity of the payload task was high. However, concurrent communication task performance was relatively unaffected by changes in overall task load when

payload task complexity was low (Figure 4.4). Conversely, when the payload task automation was more reliable, concurrent communication task performance was relatively unaffected by changes in overall task load when payload task complexity was high. However, there was a pronounced decrease in performance on the concurrent communication task attributable to increased overall task load when the complexity of the payload task was low (Figure 4.5). That being said, this three-way interaction should be interpreted with some caution given the relative scarcity of chat messages in the low condition. However, this range restriction is less of a concern since throughput was employed as a performance index instead of accuracy; the range of possible accuracy values would have been much more limited.

The results of previous studies have been mixed on whether or not automation, particularly false-alarm prone automation, affects concurrent task performance. The non-automated concurrent tasks in the present study were the UAV routing task and the chat communications task. The reliability of the payload task automation did not affect concurrent UAV-routing performance, which is consistent with the findings of Levinthal and Wickens (2006), who employed similar, less physically demanding waypoint navigation. On the other hand, performance on the concurrent chat communication task was the result of a complex three-way interaction between overall task load, the complexity of the automated payload task, and the reliability of the payload task automation. There was not a significant main effect for payload task automation reliability on concurrent chat communication performance. The present study suggests that simply concluding that decreased automation reliability and/or bias toward false alarms results in decreased concurrent task performance might be oversimplifying the problem. Operator performance on concurrent tasks was either unaffected by automation reliability or also influenced by the complexity of the automated task and operators' current task load.

However, the results of the present study were consistent with the literature when it came to performance on the automated task itself. The literature states that false-alarm prone automation affects performance on the automated task more than miss-prone automation. Indeed, the false-alarm prone, unreliable automation condition resulted in significantly decreased participant performance on the automated payload task. However, the performance decrease was probably not the result of the “cry wolf” effect (i.e., reduced compliance) simply because participants were not required to comply with the veto automation (Dixon, Wickens, & McCarley, 2007; Levinthal & Wickens, 2006; Wickens, Dixon, & Johnson, 2005). Moreover, since the false-alarm prone automation was also 4.5% less reliable, this decrease in performance could also be due to the general decrease in reliability rather than the change in automation bias, as some studies have found a significant effect of automation reliability but not bias on operator performance (e.g., Rice, 2009). However, it could also be a combination of both.

Furthermore, it is possible that a three-way interaction could have been detected for payload task performance as well, if throughput were used as the task performance measure instead of accuracy. Conceptually, throughput is a corrected rate measure giving the number of correct responses per unit of discretionary time (i.e., time used by the participant for processing and responding). Throughput is considered a more sensitive performance measure to parallel changes in speed and accuracy since the product of such changes will be magnified. In addition, throughput is known to be more stable than accuracy and response time, which tend to fluctuate across trials in sessions where the speed-accuracy trade-off phenomenon is observed (Thorne, 2006). Unfortunately, due to a data logging error, response times for the payload task were not captured and throughput could not be calculated. However, since response time on the payload task was quite range restricted (participants had only seven seconds to override the automation),

it is more likely that the throughput measure would be little (if any) more sensitive than accuracy or that accuracy would exhibit significantly more variability. Nevertheless, the failure to capture and analyze throughput values for all relevant tasks is both a limitation of the current study and a possible direction for future research.

Despite this limitation, it is nevertheless clear that task load, task environment complexity, and automation reliability all contribute to participants' UAV supervisory control performance and their interactions can, at times, be quite complex. Looking at the effects of these variables in isolation might lead to the oversimplification and misinterpretation of their effects on UAV supervisory control performance.

Furthermore, it should be noted that a significant difference in automated task performance and a significant three-way interaction between task load, automated task complexity, and automation reliability on concurrent task performance were obtained with a relatively small automation reliability manipulation, although the corresponding effect sizes were proportionally small. While Rice (2009) and Ruff, Narayanan, and Draper (2002) implemented small reliability manipulations, their studies took place within a highly controlled single-task environment and were limited in power due to a small sample size, respectively. Nevertheless, the performance variables employed in the present study were sensitive enough to pick up these differences despite the relative noise of the SCOUT testing environment, and are thus operationally viable.

5.4 Changes in Subjective Workload and Fatigue

The second goal of this study was to determine whether differences in task load, task environment complexity, and automation reliability affected participants' subjective workload and fatigue ratings, as assessed by the CSS (Ames & George, 1993). Each participant rated his or

her mean workload, maximum workload, and fatigue at the end of each of the four experimental blocks. Unfortunately, due to a data logging error, baseline fatigue and workload ratings were not captured at the beginning of the experimental scenario for half of the participants, so delta scores are not available.

Participants' subjective mean workload increased in response to both higher task load and lower payload task automation reliability. However, the increase due to both factors was small. The increase in participants' subjective mean workload due to increased task load corresponded to a mean difference of one point on the seven-point workload scale, or workload ratings of (3) "moderate activity, easily managed, considerable spare time" and (4) "busy, challenging but manageable, adequate time available" for the low and high task load blocks, respectively. The increase in participants' subjective mean workload due to decreased payload task automation reliability corresponded to less than a one-point difference on the seven-point workload scale but, when rounding was applied, corresponded to workload ratings of four and three for the low and high automation reliability blocks, respectively.

Unlike mean workload ratings, which modestly increased in response to both higher task load and decreased automation reliability, maximum workload ratings were only affected by changes in task load. Participants' maximum workload ratings increased with overall task load, with a small, approximately one-point difference in ratings of (4) "busy, challenging but manageable, adequate time available" and (5) "very busy, demanding to manage, barely enough time" between the low and high task load blocks, respectively.

However, although significant, the small magnitude of the difference in subjective workload ratings between the low and high task load blocks, and between the average and maximum workload ratings within each block, suggests that subjective workload ratings might

be influenced by additional factors other than task load. Even in the low task load blocks, participants' average and maximum subjective workload ratings were relatively high: three and four, respectively. This elevated baseline workload could be attributed to the fact that, in multitask environments like SCOUT, time-sharing between concurrent tasks or between display elements can place additional demands on working memory even when task load is relatively low. It is also possible that participants simply tried harder when task load was low and the environment seemed more manageable (Yeh & Wickens, 1988). It is also possible that one of the three additional workload components other than time loads assessed by the CSS (activity level, system demands, and safety concerns) could have contributed to the relatively elevated and stable workload ratings. Future research should incorporate a subjective measure of effort to investigate the possible dissociation between participants' task load and subjective workload ratings due to increased effort.

Unlike subjective workload, participants' subjective fatigue was not affected by changes in task load, task environment complexity, or payload task automation reliability. Participants' mean subjective fatigue rating was three ("okay; somewhat fresh") across all experimental blocks. The SCOUT experimental scenario might not have been long enough to induce fatigue.

5.5 Automation Dependence

The third goal of the present study was to determine whether differences in task load, task environment complexity, and automation reliability affected participants' automation dependence. Overall, participants agreed with the unreliable (92.5%) automation more than the reliable (97.0%) automation. Even though the automation in both conditions was relatively reliable compared to other studies, this finding was still somewhat surprising given that participants were penalized for false alarms and the unreliable automation had a liberal response

criterion (i.e., it was prone to false alarms). In fact, the finding that participants agreed with the false-alarm prone automation more often directly contradicts much of the automation reliance-compliance literature. According to the literature, in dual-task paradigms, false-alarm prone automation affects compliance and reliance (which are two behavioral manifestations of automation dependence) as much, if not more so, than miss-prone automation (Dixon, Wickens, & McCarley, 2007; Levinthal & Wickens, 2006; Wickens, Dixon, & Johnson, 2005).

However, the interpretation of the main effect of automation reliability on participants' automation dependence (i.e., their percent agreement with the payload task automation) could be misleading since there was a significant interaction between automation reliability and task load. Although participants generally tended to rely on the unreliable, false alarm-prone automation more often, their dependence on said unreliable automation was more affected by task load relative to the reliable automation. That being said, the effect size of this interaction was quite small. The log-transformed mean difference of participant reliance on the reliable (97%) automation during the high and low task load blocks was 0.050, and the log-transformed mean difference of participant reliance on the unreliable (92.5%) automation during the high and low task load blocks was 0.057.

Nevertheless, since participants agreed with the unreliable, false-alarm prone automation more often than the reliable automation, it is clear that increased automation reliability does not necessarily result in increased automation dependence. In fact, while the effect of automation reliability on automation dependence was significant, the effect size was small and participants generally agreed with the automation a large percentage of the time regardless of its reliability: 89.5% of the time in the low reliability condition versus 91.5% in the high reliability condition. Perhaps most surprisingly, there was no main effect of task load on automation dependence.

Participants tended to depend on the automation to a high degree regardless of their current task load (90.4% and 90.6% in the low and high task load blocks, respectively). This generally high automation dependence is consistent with anecdotal evidence from prior SCOUT studies that suggests that its multitask environment is sufficiently complex and time-pressured enough that operators may depend on the payload task automation irrespective of its reliability and their trust in it.

5.6 Trust in Automation

There is a substantial body of literature that postulates that automation reliability is an important factor of human use of automated systems because of its influence on operator trust, and that unreliable automation lowers operator trust, which results in underutilization of the automation (Bliss, Gilson, & Deaton, 1995; Dixon & Wickens, 2006; Dixon, Wickens, & Chang, 2005; Dixon, Wickens, & McCarley, 2007; Meyer, 2001, 2004; Rice, 2009). Indeed, while some studies have shown that trust does affect automation dependence, trust is not the sole determining factor of automation use. As Lee and See (2004) noted in their review of the automation trust and dependence literature, studies on the topic have produced many confusing and seemingly conflicting findings, which are probably at least partially attributable to the different operationalizations of trust and the confounding of its effect with other factors, such as workload, situation awareness, perceived risk, and operator self-confidence (Lee & Moray, 1994; Lee & See, 2004; Parasuraman & Riley, 1997). Lee and Moray (1992) found that, under certain conditions, operator reliance on automation did not correspond to changes in their trust. Instead, operators depended on automation when their self-reported trust in the automation exceeded their self-reported self-confidence that they could manually perform the automated task.

In fact, previous anecdotal evidence indicated that participants might not have even been cognizant of the reliability of the automation in prior SCOUT studies. Thus, the final goal of the present study was to determine whether differences in automation reliability affected participants' subjective ratings of trust in the automation. Did participants in the low and high reliability conditions notice the difference in the reliability of the payload task automation? If so, were their trust and self-confidence ratings impacted?

Participants' trust in the payload task automation and their self-confidence that they could perform the payload task manually were assessed with a four-item survey modeled after Lee and Moray's validated (1994) scale. Participants' trust ratings for the payload task automation did not significantly differ between the low and high automation reliability groups, suggesting that participants did not notice the 4.5% between-group difference in automation reliability. However, although the difference in trust ratings between the low and high automation reliability groups was not significant, participants' mean trust ratings were generally somewhat low: 6.50 and 6.39 for the low and high automation reliability groups, respectively, on a 10-point scale, with higher ratings indicating a greater degree of trust. Thus, participants in both the low and high reliability groups were somewhat skeptical of the payload task automation. Yet, despite this skepticism, their automation dependence (percent agreement with the automation) remained quite high across the board.

Since participants were skeptical of the automation's reliability, automation bias (operators' tendency to over-rely on automation) was clearly not the phenomena behind participants' high degree of automation dependence. Rather, the findings of this experiment appear to be more consistent with the related concept of automation complacency. Automation complacency is characterized by observably substandard monitoring of an automated task under

conditions of multiple-task load, when manual tasks compete with the automated task for the operator's attention (Baghieri & Jamieson, 2004; Metzger and Parasuraman, 2005; Parasuraman & Manzey, 2010; Parasuraman, Molloy, & Singh, 1993; Wickens, Dixon, Goh, & Hammer, 2005). Participants likely reallocated their attention away from the automated payload task to the manual UAV-routing and chat communication tasks to increase their overall performance, a strategy that led to high levels of automation dependence (Parasuraman & Manzey, 2010).

In fact, automation reliability did not significantly affect participants' self-reported reliance on the automation. Participants in the high automation reliability group reported relying on the automation slightly more often than participants in the low automation reliability group, a finding that would have directly contradicted the percent agreement scores if the difference between the two groups' self-reported reliance had been statistically significant. Nevertheless, although not directly comparable, participants' self-reported automation reliance ratings were relatively low compared to their percent agreement scores, suggesting that participants might have underestimated their dependence on the payload task automation. Participants' automation reliance ratings were 6.29 and 7.11 out of 10, with higher scores indicating a greater degree of reliance, for the low and high automation reliability groups, respectively.

In addition, automation reliability did not significantly affect participants' confidence that they could perform the payload task manually. However, participants' middling self-confidence ratings (mean ratings of 5.46 and 5.60 out of ten for the low and high reliability groups, respectively) suggest that they were generally not that confident that they could correctly identify potential targets without automation. It is therefore not surprising that participants reported relatively large perceived performance improvements due to the payload task automation. Perceived performance improvements for the low and high automation reliability groups were

8.28 and 7.52 out of 10, respectively. However, the perceived performance difference between the groups was not significant.

In summary, automation reliability did not significantly affect participants' subjective trust in the payload task automation, reliance on the automation, self-confidence that they could perform the payload task manually, or perceived improvement attributable to the automation. However, participants' ratings, considered in the context of their performance data, indicated that they were generally somewhat skeptical of the automation, but depended on it anyway (although they underestimated their degree of dependence). In addition, participants were only moderately confident in their ability to perform the payload task manually and generally felt that the payload task automation improved their performance. This overall picture suggests that participants were willing to unload some of their task load onto the automation, even though they were aware it was imperfect, in an effort to improve their overall performance on a resource-limited task. Perhaps most importantly, participants in the low and high reliability groups did not seem to notice the 4.5% between-group difference in automation reliability or, at the very least, their trust was not affected by this difference. While the non-significant difference in trust ratings could be attributed to the small reliability manipulation, the significant difference in between-group percent agreement scores suggests that this is not the case and self-reported trust might not reflect actual automation dependence behavior.

These results are consistent with Lee and Moray (1992), who found that operator dependence on automation did not necessarily correspond to changes in their trust alone. Rather, operators depended on automation when their trust in the automation exceeded their self-confidence that they could perform the task manually, which was indeed the case here. The concerns of Lee and Moray (1992) that their findings might not generalize to more complex

systems, intermediate LOAs, and less manageable task loads do not seem to be valid for the present study and their results do generalize, at the very least, to the SCOUT environment.

6 Conclusion

In conclusion, participant task load significantly impacted their performance on all three of the subtasks. High task load resulted in overall reduced performance on the automated payload and chat communications tasks, but improved participants' UAV routing performance, indicating that they prioritized the perceived primary task when their task load became less manageable. However, the present study gave evidence for the roles task environment complexity and automation reliability also play in UAV supervisory control performance. By considering these factors together along with task load, multiple significant interactions were revealed. While payload task complexity and the reliability of the payload task automation did

not affect participants' performance on the concurrent UAV-routing task, significant interactions were found for the concurrent communications task and for the automated payload task itself.

Participant performance on the automated payload task was negatively impacted as a result of both increased task load and decreased automation reliability. In addition, while unreliable automation with a liberal response criterion led to a more pronounced performance decrement during periods of high task load, reliable automation partially mitigated the impact of increased task load on payload task performance. In addition, participant performance on the chat communications task was affected by concurrent task characteristics. More specifically, participant performance on the communications task was the result of an interaction between their overall task load, the complexity of the concurrent payload task, and the reliability of the payload task automation.

By considering the effects of participant task load, task environment complexity, and automation reliability on UAV supervisory control performance together, the present study sought to contribute to the field's understanding of their effects on UAV supervisory control performance. By eschewing the typical reductionist approach and embracing a comparatively "noisy" design, the experimenters were able to identify multiple interactions between an operator's task load, the complexity of the task, and the reliability of their automation on their UAV supervisory control performance. In addition, the "noisy" design enabled the experimenters to test the operational viability of certain UAV supervisory control performance metrics and found that they were sensitive enough to detect the results of smaller, more realistic automation reliability and bias manipulations in a noisy testing environment.

In addition, increased task load and reduced automation reliability both resulted in modest increases in participant's subjective mean workload. Increased task load resulted in a

modest increase in participants' subjective maximum workload, but the effect of automation reliability was not significant. None of the factors—task load, task environment complexity, or automation reliability—had an appreciable effect on operator fatigue. While subjective workload ratings did not completely dissociate from the performance metrics, its relatively small variation suggests that additional factors affected participants' subjective experience of workload. Future research should incorporate a subjective measure of effort to investigate whether “trying harder” is a primary reason behind this partial dissociation.

Furthermore, though participants surprisingly agreed more frequently with the unreliable, false-alarm prone automation, they generally exhibited a high degree of automation dependence. In fact, they agreed with the automation over 90% of the time regardless of its reliability. Thus, it is clear that increased automation reliability does not necessarily result in increased automation dependence. In addition, automation reliability had no effect on operators' subjective trust or reliance on the payload task automation or perceived improvement attributable to the automation.

This generally high automation dependence, as well as the failure of participants' self-report measures to reflect the between-group difference in automation reliability despite differences in their actual dependence behavior, suggests that self-reported trust might not reflect actual automation dependence behavior. Furthermore, this observation is consistent with both the finding that operators depend on automation when their trust in the automation exceeds their self-confidence that they can manually perform the task, and anecdotal evidence from prior SCOUT studies that suggests that its multitask environment is sufficiently complex and time-pressured enough that operators may depend on the payload task automation irrespective of its reliability and their trust in it. The overall picture presented by participants' subjective and

performance data suggests that, while they were skeptical of the automation, they were willing to unload some of their task load onto it in an effort to improve their overall performance on a resource-limited task. This finding, of course, has significant implications for the implementation of any automation in UAV supervisory control environments. Due to a combination of automation complacency and conscious strategizing to maximize their performance on a resource-limited task, operators seem to generally depend on automated aids to a high degree. Thus, automation should be implemented judiciously and with careful weighting of its benefits versus the consequences of its potential failure states.

The most significant limitation of the present study was the characterization of automation dependence as percent agreement. This metric does not allow one to fully parse the effects of automation dependence and bad performance, particularly when an operator is undertasked and is primarily monitoring. For example, during a low task load block, a participant could have frequently monitored the payload task sensor feeds but failed to notice and correct any automation errors, thus not depending on the automation but rather just performing poorly. Under the current operational definition of automation dependence, percent agreement with the automation, they would be identified as exhibiting a high degree of automation dependence. It is possible that the weakness of this operational definition could have contributed to the failure to find a significant effect of task load on automation dependence.

Another limitation of the present study was the examination of only two levels of each independent variable. If a suitably sized pool of participants becomes available, future studies could examine the performance effects of a wider spectrum of task loadings and/or degrees of task complexity. Moreover, the present study only looked at two levels of reliability. Future studies could examine a wider range of reliability and/or automation bias manipulations.

In addition, the CSS, which employs single-item workload and fatigue measures, was used to obtain participants' subjective workload and fatigue ratings. Single-item measures are popular because they are quick and easy to administer. The use of single-item workload and fatigue measures was necessary in the present study due to time constraints. However, methodologists often advocate multiple-item measures because single-item measures cannot provide a reliable measure of relatively complex constructs. Moreover, while single-item measures allow for the estimation of test-retest reliability, the internal-consistency reliability of the measure cannot be determined (Loo, 2002). While the revised CSS provides the experimenter with interval level data and correlates well with other workload measures like the NASA-TLX, the implementation of a more diagnostic multiple-item workload and fatigue measure would be preferable for future studies, time permitting (Charlton, 2002).

Furthermore, concurrent SCOUT studies have utilized low-cost eye tracking to gain a richer picture of operator state, particularly during periods of low task load when traditional performance metrics are limited. However, the eye trackers are prone to data quality problems that have so far limited their viability (Appendix B). Nevertheless, NRL is working on cost-effective solutions to improve the performance of the low-cost eye trackers. Future studies could employ the eye trackers to track participant gaze data as an additional measure of automation dependence. More specifically, percent dwell time (PDT) outside of the payload task area of interest (AOI) could be used as an indicator of automation dependence during extended monitoring periods.

Finally, this study utilized a highly unique and specialized group of participants, SNAs and SNFOs. Future research could focus on comparing their SCOUT UAV supervisory control performance to other populations: gamers, private pilots, etc. At present, it is still unknown

whether there is any benefit to selecting aviators to be UAV pilots, and the answer to this question could have significant budgetary and training pipeline implications. Moreover, the experimenters have the ability to track participants' performance as they progress through primary flight training at Naval Air Station Pensacola. It is possible that SCOUT performance, or performance on particular SCOUT subtask(s), could add predictive validity to the Aviation Selection Test Battery (ASTB), the primary test used by the U.S. Navy, Marine Corps, and Coast Guard to select officer aviation program applicants (NMOTC, 2018). At present, the test still heavily weighs "stick-and-rudder" manual flight skills. SCOUT, as a supervisory control task, could add additional predictive validity as cockpit systems become more automated and multi-tasking, monitoring, and systems management become more critical to aircrew performance.

7 Significance

The findings of this study contributed to the growing body of knowledge and research on operator performance within a UAV supervisory control setting. In the future, this body of knowledge could inform personnel selection and enable the development, testing, and evaluation of future task-specific performance metrics, work support tools, and training. In particular, this study evaluated the effect of operator task load, task complexity, and automation reliability on UAV supervisory control performance using minimally intrusive or non-intrusive performance metrics (e.g., throughput in response to mission events) and subjective ratings. The most immediate goal of this research was to contribute to the development of a suite of performance metrics sensitive enough to be useful for the development and evaluation of future UAV ground control stations. In addition, understanding how task load and task complexity affect operator performance could inform mission-specific manpower requirements (e.g., a given mission may be more or

less complex, or more or less risk sensitive, and should be manned accordingly). Furthermore, it could contribute to a dynamic task allocation algorithm for distributing control of UAVs among a group of operators based on their current task load, task complexity, and mission requirements. In addition, since real-world automation is never 100% accurate, understanding how reliable and unreliable automation influences operator performance and decision making under variable levels of task load and complexity is of critical importance.

References

- Adams, M. J., Tenney, Y. J., & Pew, R. W. (1995). Situation awareness and the cognitive management of complex systems. *Human Factors*, 37(1), 85–104.
- Ajzen, I. (1988). *Attitudes, personality, and behavior*. Chicago, IL: Dorsey.
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179-211.
- Ajzen, I. & Fishbein, M. (2000). Attitudes and the attitude-behavior relation: Reasoned and automatic processes. *European Review of Social Psychology*, 11(1), 1–33.
- Ajzen, I. & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Upper Saddle River, NJ: Prentice Hall.
- Altini, A. (2014, February 1). Heart rate variability using the phone's camera. Retrieved December 30, 2014, from <http://www.marcoaltini.com/2/post/2014/01/heart-rate-variabilityusing-the-phones-camera.html>.
- Ames, L. L., & George, E. J. (1993). Revision and verification of a seven-point workload estimate scale (Report No. AFFTC-TIM-93-01). Edwards Air Force Base, CA: Air Force Flight Test Center.
- Anderson, P. W. (1972). More is different. *Science*, 177(4047), 393–396.
- Bagheri, N., & Jamieson, G. A. (2004). Considering subjective trust and monitoring behavior in assessing automation-induced “complacency.” In D. A. Vicenzi, M. Mouloua, & O. A. Hancock (Eds.), *Human performance, situation awareness, and automation: Current research and trends* (pp. 54–59). Mahwah, NJ: Erlbaum.
- Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91(2), 276–292.
- Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary & G. G. Berntson (Eds.), *Handbook of psychophysiology* (2nd ed.) (142–162). Cambridge, UK: Cambridge University Press.
- Billings, C. E. (1997). *Aviation automation: The search for a human-centered approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Billings, C. E. (1991). *Human-centered aircraft automation: A concept and guidelines* (N91-32135). Moffett Field, CA: Ames Research Center.
- Bliss, J. P. (2003). Investigation of alarm-related accidents and incidents in aviation. *The International Journal of Aviation Psychology*, 13(3), 249–268.

- Bliss, J. P., Gilson, R. D., & Deaton, J. E. (1995). Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics*, 38, 2300–3212.
- Borchers, H. W. (2016). Package ‘pracma.’ R package version 1.9.5 [Computer software]. Retrieved from <https://cran.r-project.org/web/packages/pracma/pracma.pdf>
- Bradley, J. V. (1958) Complete counterbalancing of immediate sequential effects in a Latin square design. *Journal of the American Statistical Association*, 53(282), 525–528.
- Caffier, P. P., Erdmann, U., & Ullsperger, P. (2003). Experimental evaluation of eye-blink parameters as a drowsiness measure. *European Journal of Applied Physiology*, 89(3-4), 319–325.
- Cain, B. (2007). *A review of the mental workload literature* (RTO-TR-HFM-121-Part-II). Toronto, CA: Defence Research And Development Canada Toronto.
- Calhoun, G. L., Draper, M. H., & Ruff, H. A. (2009). Effect of level of automation on unmanned aerial vehicle routing task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 53, pp. 197–201). San Antonio, TX: SAGE Publications.
- Calhoun, G., Funke, G., Matthews, G., Wohleber, R., Lin, J., Chiu, C. Y. P., & Ruff, H. (2016). *Impact of Individual Differences on Reliance Optimization* (711 HPW/RHCI). Dayton, OH: Wright-Patterson AFB.
- Chalmers, D. J. (2006). Strong and weak emergence. In P. Clayton & P. Davies (Eds.), *The Re-emergence of Emergence: The Emergentist Hypothesis from Science to Religion* (pp. 244–256). Oxford, NY: Oxford University Press.
- Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. H. (2017). Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors*, 59(3), 333–345.
- Chanda, M., DiPlacido, J., Dougherty, J., Egan, R., Kelly, J., Kingery, T., Liston, D., Mousseau, D., Nadeau, J., Rothman, T., Smith, L., Supko, M. (2010). *Proposed functional architecture and associated benefits analysis of a common ground control station for Unmanned Aircraft Systems*. Monterey, California. Naval Postgraduate School.
- Chanowitz, B., & Langer, E. J. (1981). Premature cognitive commitment. *Journal of Personality and Social Psychology*, 41(6), 1051–1063.
- Charlton, S. G. (2002). Measurement of cognitive states in test and evaluation. In S. G. Charlton & T. G. O’Brien (Eds.), *Handbook of human factors testing and evaluation* (97–126). Mahwah, NJ: Lawrence Erlbaum Associates.

- Chen, J. Y. C., Barnes, M. J., & Harper-Sciari, M. (2011). Supervisory Control of Multiple Robots: Human-Performance Issues and User-Interface Design. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 41(4), 435–454.
- Colvin, K., Funk, K., & Braune, R. (2005). Task prioritization factors: Two part-task simulator studies. *The International Journal of Aviation Psychology*, 15(4), 321–338.
- Comstock, J. R., & Arnegard, R. J. (1992). *The Multi-Attribute Task Battery for human operator workload and strategic behavior research* (NASA Tech. Memorandum No. 104174). Hampton, VA: NASA.
- Coyne, J. T., Foroughi, C., & Sibley, C. (2017). Pupil diameter and performance in a supervisory control task: A measure of effort or individual differences? *Proceedings of the Human Factors and Ergonomics Society 2017 Annual Meeting*, 61(1), 865–869.
- Coyne, J. T., & Sibley, C. M. (2015a). Impact of task load and gaze on situation awareness in unmanned aerial vehicle control. *Proceedings of the 18th International Symposium on Aviation Psychology*. Red Hook, NY: Curran Associates, Inc.
- Coyne, J., & Sibley, C. (2015b). *Supervisory Control Operations User Testbed (SCOUT): Investigating complex decision-making in a realistic virtual paradigm*. Manuscript in preparation.
- Coyne, J., & Sibley, C. (2016). Investigating the use of two low cost eye tracking systems for detecting pupillary response to changes in mental workload. *Proceedings of the 2016 Annual Meeting of the Human Factors and Ergonomics Society*. Thousand Oaks, CA: Sage Publishing.
- Coyne, J., Sibley, C., & Sherwood, S. (2016). The rise of the low-cost eye trackers: An evaluation of several new systems. Abstract presented at the 8th International Conference on Applied Human Factors and Ergonomics, July 17–21, Orlando, FL.
- Cummings, M. L., Bertucelli, L. F., Macbeth, J., & Surana, A. (2014). Task versus vehicle-based control paradigms in multiple unmanned vehicle supervision by a single operator. *IEEE Transactions on Human-Machine Systems*, 44(3), 353–361.
- Cummings, M. L., & Nehme, C. E. (2009). Modeling the impact of workload in network centric supervisory control settings. In *2nd Annual Sustaining Performance Under Stress Symposium*. College Park, MD.
- Defense Science Board (2016). Report of the Defense Science Board summer study on autonomy. Washington, DC: Department of Defense.

- Department of Defense. (2014). *Department of Defense interface standard: Joint military symbology* (MIL-STD-2525D). Department of Defense.
- Department of Defense. (2013). *Unmanned systems integrated roadmap FY2013–2038*. Washington, DC: Department of Defense.
- 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*, 48(3), 474–86.
- Dixon, S. R., Wickens, C. D., & Chang, D. (2003). Comparing quantitative model predictions to experimental data in multiple-UAV flight control. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 47(1), 104–108. Los Angeles, CA: Sage Publications.
- 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*, 49(4), 564–572.
- Dominguez, C., Vidulich, M., Vogel, E., & McMillan, G. (1994). *Situation awareness: Papers and annotated bibliography* (AL/CF-TR-1994-0085). Brooks AFB, TX: Armstrong Laboratory.
- Donmez, B., Nehme, C., & Cummings, M. L. (2010). Modeling workload impact in multiple unmanned vehicle supervisory control. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions On*, 40(6), 1180–1190.
- Durso, F. T., & Dattel, A. R. (2004). SPAM: The real-time assessment of SA. In S. Banbury & S. Tremblay (Eds.), *A cognitive approach to situation awareness: Theory and application* (Vol. 1, pp. 137–154). Hampshire, UK: Ashgate.
- Eggers, J. W. & Draper, M. H. (2006). Multi-UAV control for tactical reconnaissance and close air support missions: Operator perspectives and design challenges. In *Proceedings of the NATO RTO Human Factors and Medicine Panel Symposium HFM-3-135*, NATO RTO.
- Endsley, M. R. (1988a). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd Annual Meeting*, 97–101. Santa Monica, CA: Human Factors and Ergonomics Society.
- Endsley, M. R. (1988b). Situation Awareness Global Assessment Technique (SAGAT). *Proceedings of the IEEE National Aerospace and Electronics Conference*, 789–795. New York, NY: IEEE.

- Endsley, M. R. (2000). Theoretical underpinnings of situation awareness: A critical review. In M. R. Endsley & D. J. Garland (Eds.), *Situation analysis and measurement* (pp. 3–28). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1), 32–64.
- Endsley, M. R. and Kaber, D. M. (1999) Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, 42 (3) 462–492.
- Endsley, M. R. & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, 37(2), 381–394.
- Feldman, J. (2003). Simplicity and complexity in human concept learning. *The General Psychologist*, 38, 9–15.
- Fishbein, M. & Ajzen, I. (1975). *Belief, attitude, intention, and behavior*. Reading, MA: Addison-Wesley.
- Foroughi, C. K., Sibley, C., & Coyne, J. T. (2017). Pupil size as a measure of within-task learning. *Psychophysiology*, 54(10), 1436–1443.
- Freedberg, S. J. (2012). Too many screens: Why drones are so hard to fly, so easy to crash. *Breaking Defense*. Retrieved from <https://breakingdefense.com/2012/08/too-many-screens-why-drones-are-so-hard-to-fly-and-so-easy/>
- Funk, K., Lyall, B., Wilson, J., Vint, R., Niemczyk, M., Suroteguh, C., & Owen, G. (1999). Flight deck automation issues. *International Journal of Aviation Psychology*, 9(2), 109–123.
- Funke, G., Greenlee, E., Carter, M., Dukes, A., Brown, R., & Menke, L. (2016). Which eye tracker is right for your research? Performance evaluation of several cost variant eye trackers. *Proceedings of the 2016 Annual Meeting of the Human Factors and Ergonomics Society*. Thousand Oaks, CA: Sage Publishing.
- Gallen, Christine. (2014, November 13). Interior Cameras and Eye-tracking to Dominate Driver Monitoring Technology in Active Safety, Autonomous Driving, and Smart HMI Era, According to ABI Research | Business Wire. Retrieved December 30, 2014, from <http://www.businesswire.com/news/home/20141113005978/en/Interior-Cameras-Eye-tracking-Dominate-Driver-MonitoringTechnology#.VKMJECvF98E>.
- George, D., & Mallery, M. (2009). *SPSS for Windows step by step: A simple guide and reference, 17.0 update* (10th ed.). Boston, MA: Pearson.
- Gertler, J. (2012). *U.S. unmanned aerial systems* (CRS Report No. R42136). Washington, DC: Congressional Research Service.

- Ghosh, D., & Vogt, A. (2012). Outliers: An evaluation of methodologies. In *Joint statistical meetings* (pp. 3455-3460). San Diego, CA: American Statistical Association.
- Giese, S., Carr, D., & Chahl, J. (2013). Implications for unmanned systems research of military UAV mishap statistics. In *2013 IEEE Intelligent Vehicles Symposium (IV)* (pp. 1191–1196).
- Goldberger, A. L. (2006). Complex systems. *Proceedings of the American Thoracic Society*, 3(6), 467–471.
- Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.
- Gueorguieva, R., & Krystal, J. H. (2004). Move over ANOVA: progress in analyzing repeated-measures data and its reflection in papers published in the archives of general psychiatry. *Archives of general psychiatry*, 61(3), 310–317.
- Guerin, L., & Stroup, W. W. (2000). A simulation study to evaluate PROC MIXED analysis of repeated measures data. *Proceedings of the Conference on Applied Statistics in Agriculture*, 170–203. Manhattan, KS: New Prairie Press.
- Gurka, M. J., Edwards, L. J., & Muller, K. E. (2011). Avoiding bias in mixed model inference for fixed effects. *Statistics in medicine*, 30(22), 2696–2707.
- 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). North-Holland.
- Hebb, D. O. (1955). Drives and the CNS (conceptual nervous system). *Psychological review*, 62(4), 243.
- Hess, E. H. (1975). *The tell-tale eye: How your eyes reveal hidden thoughts and emotions*. New York, NY: Van Nostrand Reinhold, Inc.
- Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, 143(3611), 1190–1192.
- Hjortskov, N., Rissén, D., Blangsted, A. K., Fallentin, N., Lundberg, U., & Sjøgaard, K. (2004). The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal of Applied Physiology*, 92(1-2), 84–89.

- Howell, D. C. (2010). *Statistical methods for psychology* (7th ed.). Belmont, CA: Wadsworth.
- Johnson, R., Leen, M., & Goldberg, D. (2007). *Testing adaptive levels of automation (ALOA) for UAV supervisory control* (No. AFRL-HE-WP-TR-2007-0068). Dayton, OH: Air Force Research Laboratory.
- Johnson, S. (2012). *Emergence: The connected lives of ants, brains, cities, and software*. New York, NY: Scribner.
- Joslyn, C. & Rocha, L. M. (2000). Towards semiotic agent-based models of socio-technical organizations. In *Proceedings of the AI, Simulation, and Planning in High Autonomy Systems Conference* (pp. 70–79). Tucson, AZ: IEEE.
- Kaber, D. B. & Endsley, M. R. (2003). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, 5(2), 1–40.
- Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Kessel, C. J. and Wickens, C. D. (1982). The transfer of failure-detection skills between monitoring and controlling dynamic systems, *Human Factors*, 24(1), 49–60.
- Kidwell, B., Calhoun, G. L., Ruff, H. A., & Parasuraman, R. (2012). Adaptable and adaptive automation for supervisory control of multiple autonomous vehicles. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56, pp. 428–432). Boston, MA: SAGE Publications.
- Klinger, J., Kumar, R., & Hanrahan, P. (2008). Measuring the task-evoked pupillary response with a remote eye tracker. *Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications*. New York, NY: ACM.
- Langer, E. J. (1989). *Mindfulness*. Reading, MA: Addison-Wesley/Addison Wesley Longman.
- Lee, J. & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10), 1243–1270.
- Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40(1), 153–184.
- Lee, J. D. & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80.

- Levinthal, B. R., & Wickens, C. D. (2006). Management of multiple UAVs with imperfect automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(17), 1941–1944.
- Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, 83(404), 1198–1202.
- 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 & Robotic Systems*, 54(5), 795.
- Liu, Y., & Wickens, C. D. (1987). The effect of processing code, response modality and task difficulty on dual task performance and subjective workload in a manual system. In *Proceedings of the Human Factors Society Annual Meeting*, 31(7), 847–851. Los Angeles, CA: SAGE Publications.
- Loo, R. (2002). A caveat on using single-item versus multiple-item scales. *Journal of Managerial Psychology*, 17(1), 68–75.
- Luximon, A., & Goonetilleke, R. S. (2001). Simplified subjective workload assessment technique. *Ergonomics*, 44(3), 229–243.
- Macmillan, N. A. (2002). Signal detection theory. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology (3rd ed.): Volume 4, methodology in experimental psychology* (43–90). New York, NY: John Wiley & Sons, Inc.
- Magezi, D. A. (2015). Linear mixed-effects models for within-participant psychology experiments: An introductory tutorial and free, graphical user interface (LMMgui). *Frontiers in Psychology*, 6, 2.
- Maltz, M. & Shinar, D. (2003). New alternative methods of analyzing human behavior in cued target acquisition. *Human Factors*, 45(2), 281–295.
- Marshall, S. P. (2007). Identifying cognitive state from eye metrics. *Aviation, space, and environmental medicine*, 78(Supplement 1), B165-B175.
- May, P., Molloy, R., & Parasuraman, R. (1993, October). *Effects of automation reliability and failure rate on monitoring performance in a multitask environment*. Paper presented at the Annual Meeting of the Human Factors Society, Seattle, WA.
- McFadden, S. M., Giesbrecht, B. L., & Gula, C. A. (1998). Use of an automatic tracker as a function of its reliability. *Ergonomics*, 41(4), 512–536.
- Mekdeci, B., & Cummings, M. L. (2009). Modeling Multiple Human Operators in the Supervisory Control of Heterogeneous Unmanned Vehicles. In *Proceedings of the 9th*

- Workshop on Performance Metrics for Intelligent Systems* (pp. 1–8). New York, NY: ACM.
- Metzger, U., & Parasuraman, R. (2005). Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. *Human Factors, 47*, 35–49.
- Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors, 43*(4), 563–572.
- Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors, 46*(2), 196–204.
- Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied, 6*(1), 44–58.
- NATO. (2012). *Standard Interfaces of UAV Control System (UCS) for NATO UAV Interoperability* (No. STANAG 4586 (Edition 3)). Brussels, Belgium: NATO Standardization Agency.
- Navy Medicine Operational Training Center. (2018). Retrieved March 12, 2018, from <https://www.med.navy.mil/sites/nmotc/nami/Pages/ASTBFrequentlyAskedQuestions.aspx>
- Nehme, C. E. (2009). *Modeling human supervisory control in heterogeneous unmanned vehicle systems*. Cambridge, MA: Massachusetts Institute of Technology.
- Nevin, J. A. (1969). Signal detection theory and operant behavior: A review of David M. Green and John A. Swets' signal detection theory and psychophysics. *Journal of the Experimental Analysis of Behavior, 12*(3), 475–480.
- Office of Naval Research. (2015). *Naval S&T Strategy*. Arlington, VA: ONR.
- Office of the Secretary of Defense (2012) Unmanned aircraft systems ground control station human-machine interface: Development and standardization guide, Office of the Under Secretary of Defense, Washington, DC.
- Oliver, R., Bjoertomt, O., Greenwood, R., & Rothwell, J. (2008). 'Noisy patients'—can signal detection theory help?. *Nature Reviews Neurology, 4*(6), 306.

- Ooms, K., Dupont, L., Lapon, L., & Popelka, S. (2015). Accuracy and precision of fixation locations recorded with the low-cost Eye Tribe tracker in different experimental set-ups. *Journal of Eye Movement Research*, 8, 1–24.
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381–410.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics- Part A: Systems and Humans*, 30(3), 286–297.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making*, 2(2), 140–160.
- Pariès, J. (2006). *Complexity, emergence, resilience...*. In D. Woods & N. Leveson (Eds.), *Resilience Engineering* (pp. 43–53). Burlington, VT: Ashgate Publishing Company.
- Pearson, R. K. (2002). Outliers in process modeling and identification. *IEEE Transactions on control systems technology*, 10(1), 55–63.
- Poole, A. & Ball, L. J. (2005). Eye tracking in human-computer interaction and usability research: Current status and future prospects. In C. Ghaoui (Ed.), *Encyclopedia of Human Computer Interaction* (pp. 211–219). Hershey, PA: Idea Group Reference.
- Ratwani, R. M., McCurry, J. M., & Trafton, J. G. (2010). Single Operator, Multiple Robots: An Eye Movement Based Theoretic Model of Operator Situation Awareness. In *Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction* (pp. 235–242). Piscataway, NJ, USA: IEEE Press.
- Rayner, K. & Pollatsek, A. (1989). *The psychology of reading*. Englewood Cliffs, NJ: Prentice Hall.
- Rice, S. (2009). Examining single- and multiple-process theories of trust in automation. *The Journal of General Psychology*, 136(3), 303–322.
- Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task, *Human Factors*, 49(1), 76–87.

- 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. In *Proceedings of the Human Performance, Situation Awareness, and Automation Technology Conference*, 218–222.
- 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.
- Samn, S. W. & Perelli, L. P. (1982). *Estimating aircrew fatigue: A technique with application to airlift operations* (ADA125319). Brooks Air Force Base, TX: USAF School of Aerospace Medicine.
- Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical methods in medical research*, 8(1), 3–15.
- Schnell, T., McLean, A., Neville, K., Rediger, S., & Sherwood, S. (2014). *Avionics and simulation design guidelines for the VCR LAD* (Technical Report). Cedar Rapids, IA: Rockwell Collins.
- Seltman, H. J. (2018). *Experimental Design and Analysis*. Retrieved from <http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf>
- Sheridan, T. B. (2012). Human supervisory control. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics, Fourth Edition* (pp. 990–1015). Hoboken, NJ: John Wiley & Sons, Inc.
- Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. Cambridge, MA: Man-Machine Systems Laboratory, Department of Mechanical Engineering, MIT.
- Sibley, C., Coyne, J., Avvari, G. V., Mishra, M., & Pattipati, K. R. (2016). Supporting multi-objective decision making within a supervisory control environment. Paper to be presented at 18th International Conference on Human-Computer Interaction (HCI 2016), July 17–22, Toronto, Canada.
- Sibley, C. M., Coyne, J. T., & Baldwin, C. (2011). Pupil dilation as an index of learning. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 55, 237–241. Thousand Oaks, CA: SAGE Publications.
- Sibley, C., Coyne, J., & Morrison, J. (2015). Research considerations for managing future unmanned systems. In *Papers from the 2015 AAAI Spring Symposium on Foundations of Autonomy and Its (Cyber) Threats: From Individuals to Interdependence*. Palo Alto, CA: IAAAI Press.

- Sibley, C., Coyne, J. T., & Sherwood, S. (2016). Research considerations and tools for evaluating human-automation interaction with future unmanned systems. In W. F. Lawless, R. Mittu, D. Sofge, & S. Russell (Eds.), *Autonomy and artificial intelligence: A threat or savior?* (pp. 157–178). Cham, Switzerland: Springer.
- Sibley, C., Coyne, J., & Thomas, J. (2016). Demonstrating the Supervisory Control Operations User Testbed (SCOUT). *Proceedings of the 2016 Annual Meeting of the Human Factors and Ergonomics Society*. Thousand Oaks, CA: Sage Publishing.
- Tabachnick, B. G. & Fidell, L. S. (2013). *Using multivariate statistics*. Upper Saddle River, NJ: Pearson Education, Inc.
- Teixeira, A., Matos, A., Souto, A., & Antunes, L. (2011). Entropy measures vs. Kolmogorov complexity. *Entropy*, 13(3), 595–611.
- Toma, T. (2015). Modeling task prioritization behaviors in a time-pressured multitasking environment (Doctoral dissertation, Oregon State University, Corvallis, Oregon). Retrieved from http://ir.library.oregonstate.edu/concern/graduate_thesis_or_dissertations/rv042z625
- Tsai, Y. F., Viirre, E., Strychacz, C., Chase, B., & Jung, T. P. (2007). Task performance and eye activity: predicting behavior relating to cognitive workload. *Aviation, space, and environmental medicine*, 78(Supplement 1), B176-B185.
- Uhlarik, J. & Comerford, D.A. (2002). *A review of situation awareness literature relevant to pilot surveillance functions* (Report No. DOT/FAA/AM-02/3). Washington, DC: Federal Aviation Administration.
- Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, 37(3), 498–505.
- U.S. Navy. (2018). *Manual of the medical department U.S. Navy* (NAVMED P-117). Washington, DC: US Government Printing Office.
- Wang, J. T. (2011). Pupil dilation and eye tracking. In M. Schulte-Mecklenbeck, A. Kühberger, & R. Ranyard (Eds.), *A handbook of process tracing methods for decision research* (pp. 185–204). New York, NY: Psychology Press.
- Wegener, I. (1987). *The complexity of Boolean functions*. Chichester, UK: John Wiley & Sons, Ltd.
- Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, 34(1–2), 28–35.

- Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & R. Davies (Eds.), *Varieties of attention* (pp. 63–101). New York: Academic Press.
- Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, 3(2), 159–177.
- Wickens, C. D. (2008). Multiple resources and mental workload. *Human factors*, 50(3), 449–455.
- Wickens, C. D., Dixon, S. R., & Johnson, N. R. (2005). *UAV automation: Influence of task priorities and automation imperfection in a difficult surveillance task* (AHFD-05-20/MAAD-05-6). Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Williams, K. W. (2004). *A summary of unmanned aircraft accident/incident data: Human factors implications* (No. DOT/FAA/AM-04/24). Federal Aviation Administration Oklahoma City, OK.
- Winter, B. (2013). *A very basic tutorial for performing linear mixed effects analyses (tutorial 2)* [PDF document]. Retrieved from http://www.bodowinter.com/tutorial/bw_LME_tutorial.pdf
- Woods, D. D., & Roth, E. M. (1988). Cognitive systems engineering. In M. Helander (Ed.), *Handbook of human-computer interaction* (pp. 3–43). Amsterdam, NL: North-Holland.
- Yeh, Y. Y., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors*, 30(1), 111–120.
- Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of comparative neurology*, 18(5), 459–482.
- Young, L. R. A. (1969). On adaptive manual control. *Ergonomics*, 12(4), 635–657.
- Zuniga, J., McCurry, M., & Trafton, J. G. (2014). A process model of trust in automation: A signal detection theory based approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1), 827–831. Los Angeles, CA: Sage Publications.

Appendix A: Supervisory Control Operations User Testbed (SCOUT) Operation

A.1 Overview

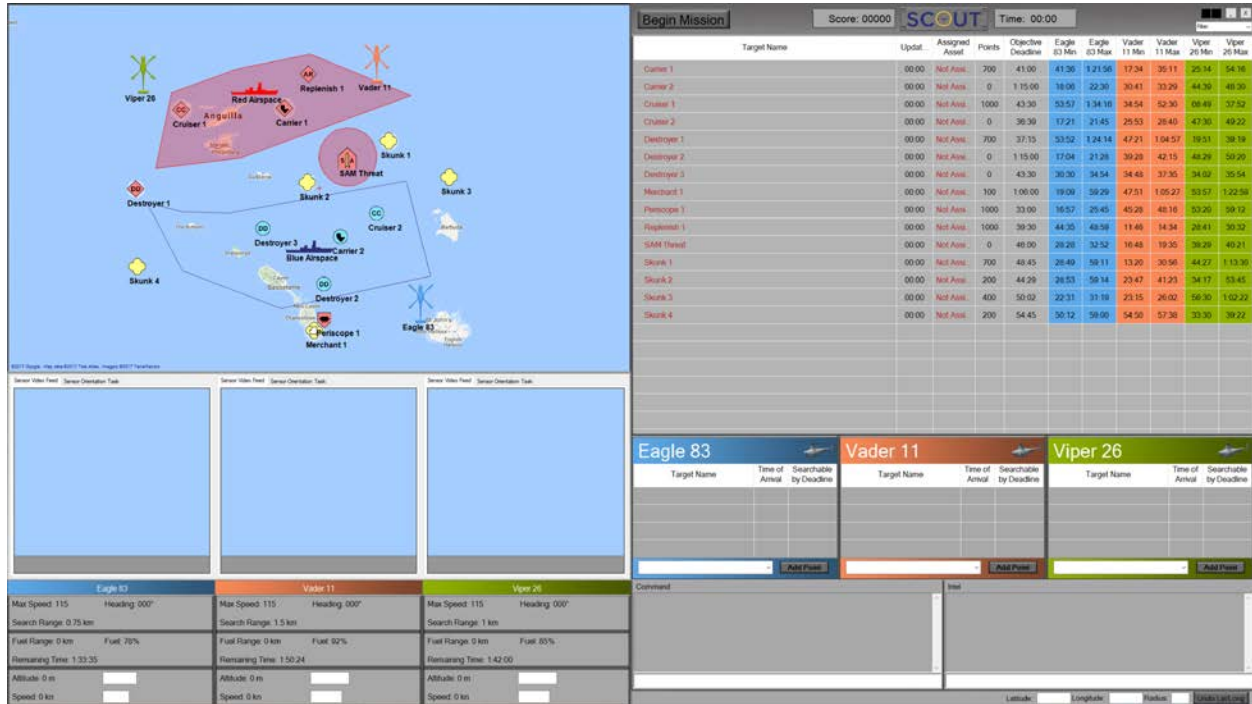


Figure A.1. The single-screen SCOUT GUI variant.

The Supervisory Control Operations User Testbed (SCOUT), developed by the Naval Research Laboratory (NRL), is a realistic simulation environment for assessing single operator performance monitoring multiple unmanned aerial vehicles (UAVs). It is designed to replicate the complexity, noise, and uncertainty associated with military UAV control. In addition, it includes tasks representative of current operators' primary roles: route planning, airspace management, communication, and monitoring (Coyne & Sibley, 2015b) (Figure A.1).

During a SCOUT mission, participants manage three heterogeneous helicopter UAVs. To meet mission goals, they must decide how to best allocate the UAVs to locate targets while simultaneously completing several subtasks, including maintaining communication with command and intelligence personnel via chat, updating UAV parameters, and monitoring their

sensor feeds and airspace. Points are assigned to various actions based on their mission priority and the goal is to obtain as many points as possible.

A.2 Demographics and Initial Setup

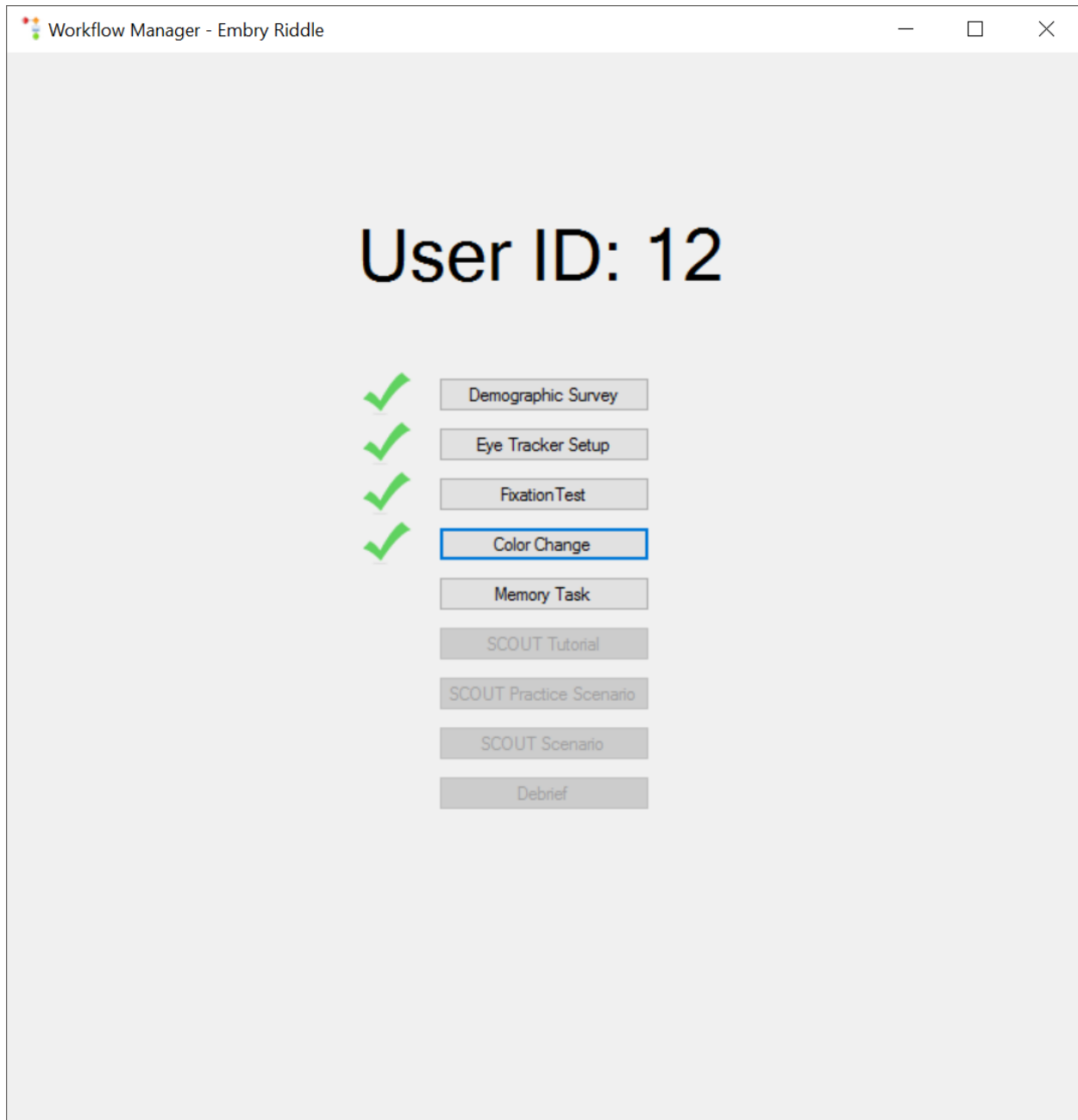


Figure A.2. Workflow Manager (Embry-Riddle variant). Completed tasks are indicated by a green checkmark. Subsequent tasks are grayed out until the participant completes all prior tasks.

A.2.1 Workflow manager. After completing all necessary informed consent and data release documentation, participants begin the SCOUT setup process through the Embry-Riddle variant of the Workflow Manager (Figure A.2.). The workflow manager guides them step-by-step through a basic demographic survey, a SCOUT tutorial, and a practice scenario before they engage in the main experimental mission. If the experiment were to utilize eye tracking, the workflow manager also includes options for eye tracker set-up, an eye tracker fixation accuracy test, and baseline tests to gauge pupil diameter change in response to changes in workload and screen luminance that can be enabled.

Survey 1.0

- □ ×

Subject ID: 12

Survey Questions:

Age: <input type="text" value="0"/>
Gender <input type="radio"/> Male <input type="radio"/> Female
Approximately how many hours a month do you spend playing video/computer games: <input type="text" value="0"/>
What is your gaming skill level? <input type="radio"/> Novice <input type="radio"/> Intermediate <input type="radio"/> Expert
Do you wear glasses: <input type="radio"/> Yes <input type="radio"/> No
Do you wear contact lenses: <input type="radio"/> Yes <input type="radio"/> No
Are you left or right-handed: <input type="radio"/> Left <input type="radio"/> Right <input type="radio"/> Ambidextrous
Are you left or right-eye dominant: <input type="radio"/> Left <input type="radio"/> Right
If you know it, please provide your visual acuity (e.g. 20/20, 20/30): <input type="text"/>
Do you have any operational experience with an Unmanned Vehicle: <input type="radio"/> Yes <input type="radio"/> No If Yes please Answer the following questions: What was your role, e.g. Pilot, Payload, Mission Commander, Other: <input type="text"/> Approximately how many years experience do you have with UAVs: <input type="text"/>
Have you ever had a pilot's license: <input type="radio"/> Yes <input type="radio"/> No
Do you currently have a pilot's license: <input type="radio"/> Yes <input type="radio"/> No

Figure A.3. Demographic survey.

A.2.2. Demographic survey. After opening the Workflow Manager, the first form each participant will fill out is a basic demographic survey (Figure A.3). The survey requests information about participant gender, gaming experience, vision, hand and eye dominance, and manned and unmanned aircraft piloting experience. The experimenter will show participants how to determine their dominant eye at the beginning of the session. More specifically, they will be instructed to extend their arms out in front of them and form a triangle with the thumbs and fingers of both their hands. They will then be asked to focus on a distant object (at least 10 ft. away) through the triangle with both eyes open. Next, they will be prompted to close each eye, one at a time, while keeping their focus on the distant object. The eye that is open when the distant object remains centered within the triangle is their dominant eye. If the distant object moves out of the frame of the triangle, the eye that is currently open is their non-dominant eye.

A.3 Mission Training

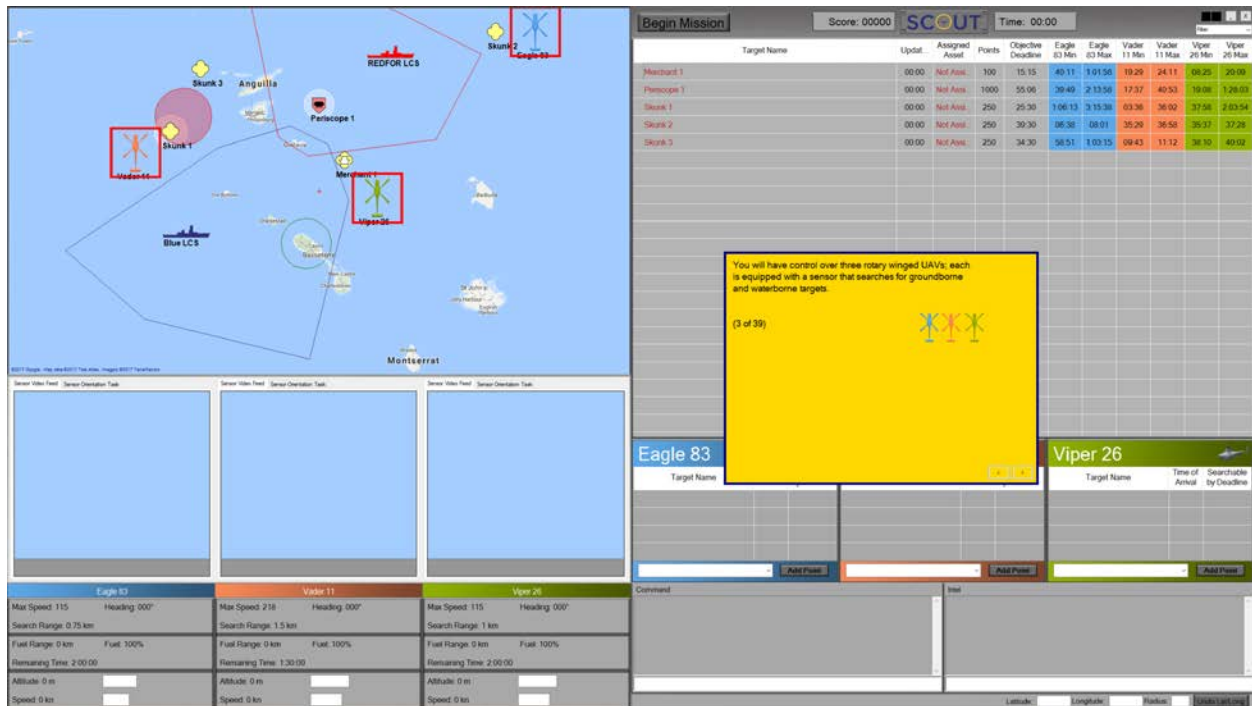


Figure A.4. SCOUT training walkthrough. The walkthrough is generalized for use with multiple concurrent studies. For this reason, Vader 11 appears as a MQ-8B Fire Scout during training.

Once baseline testing is completed, participants are trained on the operation of SCOUT. Training is conducted via a self-paced walkthrough of the test bed, which should take approximately 30 minutes to complete (Figure A.4). The training employs a series of 39 slides to instruct participants on the operation of various aspects of the SCOUT interface. When appropriate, visual indicators are used to either direct participants' attention to specific areas of interest on the GUI or to guide them to complete various actions.

After the training walkthrough, participants complete an approximately 13:45 practice mission that includes one 45-second workload and fatigue probe. Participants are encouraged to ask the experimenter for clarification on any aspect of the SCOUT test bed and its operation, if needed, before continuing on to the experimental scenario.

A.4 Mission Components and Planning

A.4.1. UAV characteristics and capabilities.

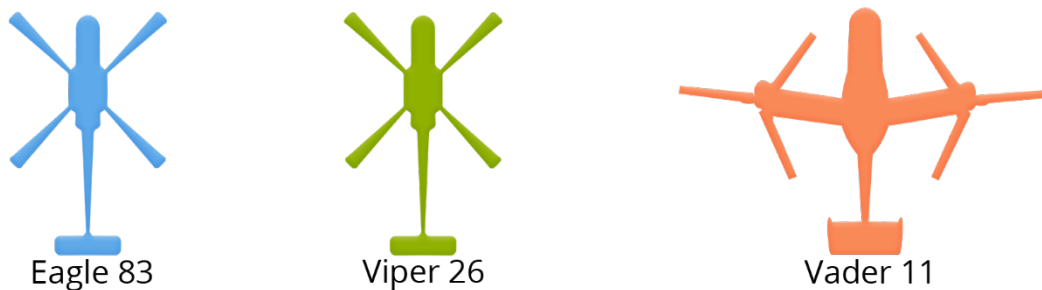


Figure A.5. UAV ownership icons.

During a SCOUT mission, participants have control over three heterogeneous helicopter UAVs: Eagle 83, Viper 26, and Vader 11. These UAVs are differentiated on the moving map display by color and name label (Figure A.5).

Table A.1

UAV Capabilities for Simulated SCOUT ISR Mission

UAV CALL SIGN	DESIGNATION	CRUISE SPEED (KTS)	MAXIMUM SPEED (KTS)	PAYLOAD TYPE	PAYLOAD RANGE (KM)
EAGLE 83	MQ-8B Fire Scout	80	85	EO/IR	1.5
VIPER 26	MQ-8C Fire Scout	115	135	EO/IR	0.75
VADER 11	RQ-10A Kestrel	210	225	EO/IR	1.2














Eagle 83 is an MQ-8B Fire Scout, Viper 26 is a MQ-8C Fire Scout, and Vader 11 is a fictional unmanned transverse rotor helicopter called the “RQ-10A Kestrel.” Fire Scouts are used for reconnaissance though, as multi-mission aircraft, they can also be used for ground support. The primary purpose of the fictional RQ-10A is reconnaissance. These three heterogeneous UAVs will be assigned to participants in a simulated SCOUT Intelligence Surveillance and Reconnaissance (ISR) mission. The payload of each of these UAVs includes a simulated electro-optical (EO/IR) sensor that will be used to search for targets on the ground. See Table A.1 for a summary of UAV capabilities.

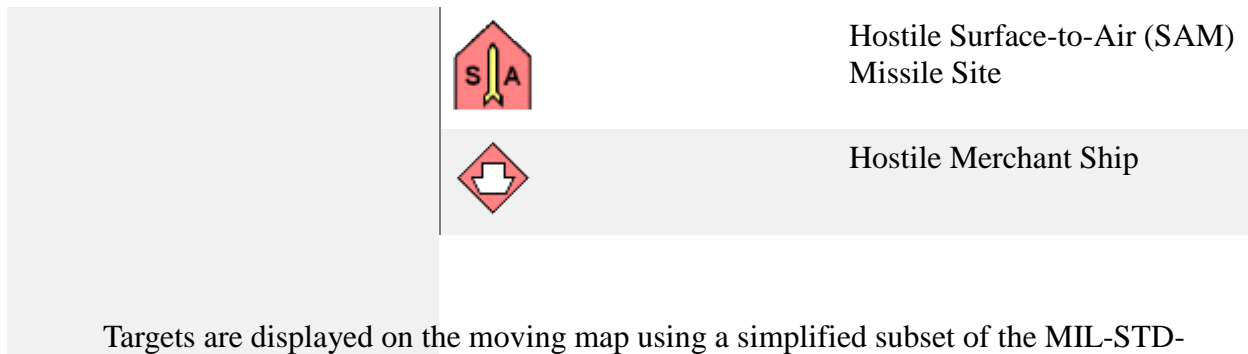
The fuel status of each UAV is displayed below its sensor feed. The participant will be asked to provide information about various fuel states through chat communication, but is informed during training that each UAV has enough fuel to complete each mission. The fuel will burn normally, but will automatically refill once empty. Participants will not have to refuel at any point or worry about running out of fuel.

A.4.2. Target characteristics.

Table A.2

Select Symbols from MIL-STD-2525D Symbol Set

	SYMBOL	MEANING
FRIENDLY		Friendly Carrier
		Friendly Cruiser
		Friendly Destroyer
		Friendly Replenishment Ship
NEUTRAL		Neutral Merchant Ship
UNKNOWN		Unknown Merchant Ship
		Unknown Surfaced Submarine
		Unknown Surface Target
HOSTILE		Hostile Carrier
		Hostile Cruiser
		Hostile Destroyer
		Hostile Surfaced Submarine
		Hostile Replenishment Ship



Targets are displayed on the moving map using a simplified subset of the MIL-STD-2525D symbol set (Table A.2) (DoD, 2014). The color of the symbols gives the operator a cursory idea of the allegiance of the target; blue is friendly, green is neutral, yellow is unknown, and red is hostile. Beyond the color, the symbol itself informs the operator of target type and mission priority. Mission priority is also reflected by a target's point value. For example, a hostile surfaced submarine will be worth far more points than a neutral merchant ship. Generally, red icons indicate targets of the greatest mission priority and point value followed by yellow icons, green icons, and blue icons.



Figure A.6. Visualization of search area size.

The precise latitude and longitude of each target is uncertain. The search area of each target is represented on the map by a surrounding white circle (Figure A.6). Some search areas are quite large, but others are so small that they are hidden by the MIL-STD-2525D icon. Each target exists somewhere within its search area, but the time that it will take to completely search increases as a function of the size of the search area.

A.4.3. Route planning. The SCOUT scenario begins with a planning period, which is untimed but should take participants approximately five to ten minutes to complete. During the planning period, participants decide where to send their three assigned UAVs to search five mission-relevant target areas. Each mission-relevant target has an associated point value, which is indicative of its mission priority. There are additional targets on the map at the beginning of the scenario, but they are not relevant to the UAV mission and are thus not assigned a point value (e.g., friendly forces).

Target Name	Updat...	Assigned Asset	Points	Objective Deadline	Eagle 83 Min	Eagle 83 Max	Vader 11 Min	Vader 11 Max	Viper 26 Min	Viper 26 Max
Carrier 1	00:00	Not Assi...	700	41:00	41:36	1:21:56	17:34	35:11	25:14	54:16
Carrier 2	00:00	Not Assi...	0	1:15:00	18:06	22:30	30:41	33:29	44:39	46:30
Cruiser 1	00:00	Not Assi...	1000	43:30	53:57	1:34:16	34:54	52:30	08:49	37:52
Cruiser 2	00:00	Not Assi...	0	36:39	17:21	21:45	25:53	28:40	47:30	49:22
Destroyer 1	00:00	Not Assi...	700	37:15	53:52	1:24:14	47:21	1:04:57	19:51	39:19
Destroyer 2	00:00	Not Assi...	0	1:15:00	17:04	21:28	39:28	42:15	48:29	50:20
Destroyer 3	00:00	Not Assi...	0	43:30	30:30	34:54	34:48	37:35	34:02	35:54
Merchant 1	00:00	Not Assi...	100	1:06:00	19:09	59:29	47:51	1:05:27	53:57	1:22:59
Periscope 1	00:00	Not Assi...	1000	33:00	16:57	25:45	45:28	48:16	53:20	59:12
Replenish 1	00:00	Not Assi...	1000	39:30	44:35	48:59	11:46	14:34	28:41	30:32
SAM Threat	00:00	Not Assi...	0	46:00	28:28	32:52	16:48	19:35	38:29	40:21
Skunk 1	00:00	Not Assi...	700	48:45	28:49	59:11	13:20	30:56	44:27	1:13:30
Skunk 2	00:00	Not Assi...	200	44:29	28:53	59:14	23:47	41:23	34:17	53:45
Skunk 3	00:00	Not Assi...	400	50:02	22:31	31:19	23:15	26:02	56:30	1:02:22
Skunk 4	00:00	Not Assi...	200	54:45	50:12	59:00	54:50	57:38	33:30	39:22

Figure A.7. The Target Information table.

The Target Information table displays the point value and deadline for each target (Figure A.7). The target deadline represents the point at which the intelligence (i.e., the approximate location of the target) is no longer valid. To receive points, the participant must locate a target by its deadline.

In addition to displaying the point value and deadline of each target, the Target Information table shows the “min.” and “max.” search times for each UAV. The min. time is the clock time at which the UAV will arrive at the target search area. The max. time is the clock time

at which the UAV will have searched 100% of the target search area. A participant can click on the headers in the Target Information table to sort targets by their points and deadlines. These times will update during gameplay as the UAV is moving.

Depending on where the target is located within its search area, it could take a participant any amount of time between the min. and max. clock times to find it. Since Vader 11 has the fastest cruise speed, it will reach target search areas faster than Eagle 83 and Viper 26. Although Eagle 83 is the slowest UAV and will take the longest to arrive at a given search area, it has the greatest sensor range so it will be able to cover the target search area faster once it arrives. If the deadline of a target precedes its max. search time, there is a chance that the target might not be located. For example, Viper 26 will finish searching 100% of Destroyer 1 at clock time 39:19. However, the deadline of Destroyer 1 is 37:15. Therefore, unless Destroyer 1 was located within the percentage of the search area covered by 37:15, it will not be located.

Eagle 83			Vader 11			Viper 26		
Target Name	Time of Arrival	Searchable by Deadline	Target Name	Time of Arrival	Searchable by Deadline	Target Name	Time of Arrival	Searchable by Deadline
Periscope 1	18:57	100%	Waypoint 0	08:00	0%	Destroyer 1	19:51	88%
Merchant 1	28:28	93%	Skunk 1	18:22	100%	Skunk 4	50:55	65%
			Skunk 3	48:21	61%			
			Waypoint 1	1:07:34	0%			

Figure A.8. UAV route builder boxes.

Once the participant makes a cursory decision on where they want to send their UAVs, they can assign targets to the vehicles by dragging the targets from the Target Information table to the UAV Route Builder Boxes (Figure A.8). Unlike the min. times in the Target Information table, which display the time of arrival were a UAV immediately directed to a target, the UAV Route Builder boxes display the time of arrival at each target based on the UAV’s current route plan. For example, per the route assigned to Vader 11, the UAV will pass through Waypoint 0

before arriving at Skunk 1 at 18:22. According to the Target Information table, Vader 11 will search 100% of Skunk 1 by 30:56. Since the deadline of Skunk 1 is 48:45, it will be able to complete the search well before the target deadline. The Searchable by Deadline column in Vader 11's Route Builder Box confirms that the UAV can search 100% of the target before its deadline. After Vader 11 searches 100% of Skunk 1, it will move onto the next target, Skunk 3. It will arrive at Skunk 3 at 48:21 and will cover 61% of the search area before the target deadline at 50:02. If Skunk 3 is not located by 50:02, Vader 11 will automatically break off the search and move on to Waypoint 1. If no subsequent target is assigned, Vader 11 will loiter at Skunk 3's former location.

The arrival times in the Route Builder Box are subject to change since the arrival times for targets with one or more preceding stops assume the UAV will search all prior targets until their deadlines are up or until 100% of their search area is covered, whichever comes first. Since a target might be located after only a small percentage of its search area is covered, the estimated arrival times could be later than is achievable. Participants should keep this in mind as they plan.

During gameplay, new targets will become available for the participant to pursue. The participant should check the Target Information table frequently so they do not miss any new targets and opportunities to score points. When a new target appears, the Target Information table will inform them how quickly each of their UAVs can get to that target area if they were to go immediately there. Participants may replan accordingly, possibly changing their original route.

The Route Builder Boxes are also useful tools for participants to refine their route plan. Participants can delete targets (by hitting the "Delete" key) or change the target order (by using the arrows on their keyboard) to see how it affects the time of arrival and Searchable by Deadline

percentages for each assigned target. This will help them determine their optimal route, which will vary according to how much risk they are willing to take pursuing different targets.

A.4.4. Route automation limitations. Although the route builders are valuable tools, participants are advised during training that SCOUT UAV routing automation is imperfect. It is thus important for the participant to monitor their route plan and UAVs to make sure they are behaving as expected. Common automation errors include: (1) a UAV drawing a path to a target that has expired, (2) a UAV continuing to loiter at an expired or located target (this usually occurs if the target was the last in the participant's route plan), and (3) a target's estimated time of arrival not updating in the route builder if the participant changes target order in the Route Builder mid-search. If any of these errors occur, or if route paths for a UAV appear odd for any reason, the participant can quickly reset the route by deleting and adding targets back into the Route Builder box.

A.4.5. Waypoints.



Figure A.9. Waypoint drop-down menu.





SCOUT includes route automation that will send each UAV to its next assigned target when its current target is either located or expires, whichever comes first. If a subsequent target is not assigned, the UAV will loiter at its current position. Nevertheless, there may be instances in which a participant might want to modify the course of a UAV; they can accomplish this by dropping a waypoint. To drop a waypoint, the participant must click on the map where they wish

to place the point and select “Add Point” (Figure A.9). They can then select their new waypoint from the drop-down menu at the bottom of the appropriate UAV Route Builder Box and click “Add Point.”

A.4.6. Restricted operating zones (ROZs).

Table A.3

Restricted operating zone (ROZ) types

ROZ TYPE				
PERMISSION REQUIRED	No	Yes	Yes	Yes
ACCESS GRANTED	Always	Always	Always	Never

The areas that are outlined in red, green, and blue on the map are restricted operating zones (ROZs). Any time that a UAV path crosses through a red or green-outlined ROZ, the participant must request access for the UAV to fly through that area. If a UAV’s path intersects a red-filled ROZ boundary, the participant must navigate around the area by dropping waypoints since access will never be granted to red-filled ROZs. Participants do not need to request access to blue outlined areas, as they represent friendly airspace (Table A.3).

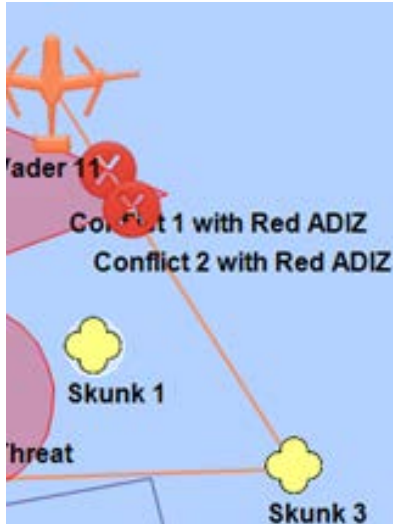


Figure A.10. ROZ incursion point (no access granted).

Participants may request access to ROZs during the planning phase; they do not have to wait until the game is in play. If a UAV will intersect a ROZ on the way to its next target, a ROZ incursion point icon will appear at the intersection of its flight path and the ROZ boundary. If this icon is a red circle with an “X” cutout, the participant must request access to enter the ROZ (Figure A.10). If a UAV enters a restricted zone without prior approval, the participant will lose a large number of points.

Table A.4

ROZ incursion point icons

ICON			
ACCESS	Not requested / Denied	Pending	Granted

Participants can request access to enter a ROZ by clicking on the ROZ incursion point icon and then clicking “Request” in the resulting pop-up box. The icon will then turn yellow to signify that the request has been submitted and is pending approval. If the icon turns green, permission to enter the ROZ was obtained. If the icon returns to red, the participant does not

have permission to enter the ROZ and must navigate around it by dropping waypoints (Table A.4). Note that approval or denial is only given during gameplay, so if a participant requests ROZ access during the planning phase, they must wait until the mission clock starts to see whether their request has been approved or denied. The ROZ incursion point icon will remain yellow until that point. If a participant is refused access to a ROZ, they will need to maneuver around that airspace to avoid losing points.

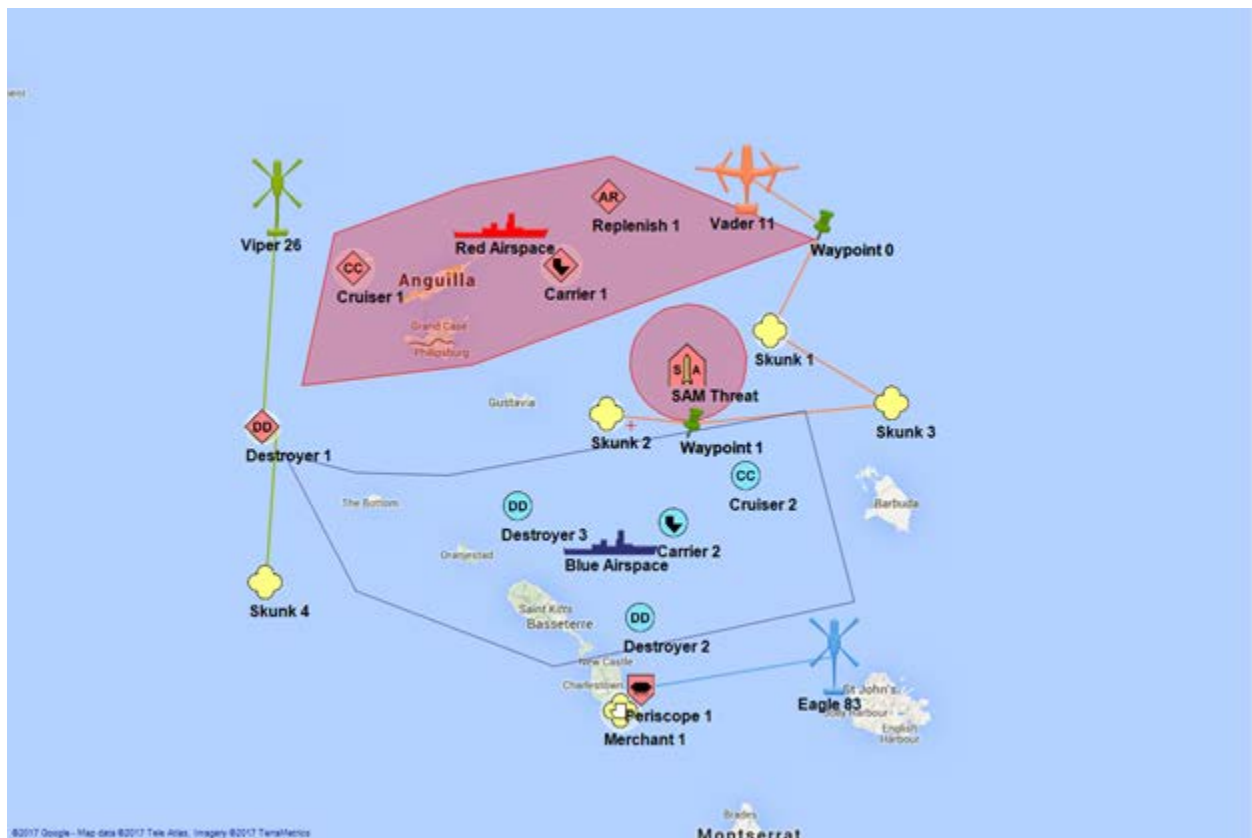


Figure A.11. Moving map display after route planning.

In summary, participants can use a combination of information from the moving map display, the Target Information table, and the UAV Route Builder boxes to formulate the best plan for sending their UAVs to targets and maximizing their points. Figure A.11 shows the moving map display at the end of a hypothetical planning period.

A.5 Gameplay

Once the participant is satisfied with their route plan, they may click the “Begin” button to start the mission. During the mission, the participant will be required to monitor and make status updates to their assigned UAVs, maintain communication with command and intelligence personnel, and locate the targets on the sensor feeds when they are within search range.

Additionally, participants will likely need to adjust their mission plan as new targets of interest appear.

A.5.1. Communication.

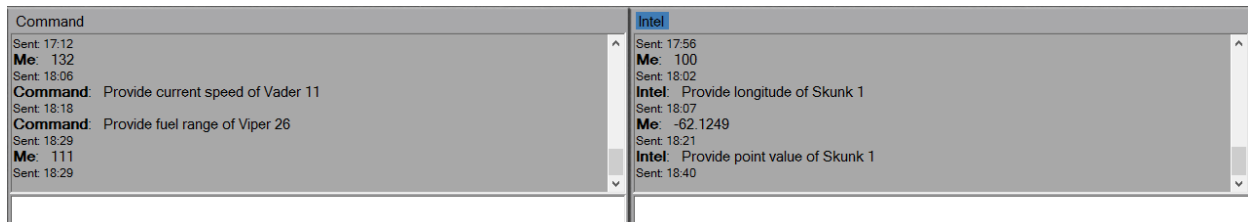


Figure A.12. Communication boxes.

Within the two communication boxes, participants receive messages from Command and Intelligence (Figure A.12). In the left box, participants receive messages from Command. Messages from Command will prompt them to make updates to vehicle statuses or to provide vehicle-related information over chat. In the right box, participants receive messages from Intelligence. Intelligence messages either query the participant for target-related information or provide information to help them locate targets faster, such as updates on search radius size or, occasionally, providing exact target coordinates.

Table A.5

Command chat messages

<i>Message</i>	<i>Required Action(s)</i>	<i>Points</i>
<i>Provide heading of [randomized UAV]</i>	Estimate heading of asset or click on specified asset in the moving map display and note the heading listed the pop-up box. Enter value into Command chat box and hit “enter”	25
<i>Provide current speed of [randomized UAV]</i>	Enter speed, no unit required, into Command chat box. Hit “enter”	25
<i>Provide fuel range of [randomized UAV]</i>	Enter speed, no unit required, into Command chat box. Hit “enter”	25
<i>Provide remaining fuel (%) for [randomized UAV]</i>	Enter remaining fuel percent, no unit required, into Command chat box. Hit “enter”	25
<i>Increase altitude of [randomized UAV] by [random integer between 1 and 150, but must be between 100 and service ceiling of UAV]</i>	Add specified integer to current altitude. Input new value into speed entry field, no unit required, and hit “enter”	25
<i>Decrease altitude of [UAV] by [random integer between 1 and 150, but must be between 100 and service ceiling of UAV]</i>	Subtract specified integer from current altitude. Input new value into speed entry field, no unit required, and hit “enter”	25
<i>Aim [randomized UAV]'s sensor at [nearby target]</i>	Aim UAV’s sensor at target within the UAV’s Sensor Orientation Tab. Detailed instructions follow in Section A.5.2	100

Table A.6

Intelligence chat messages

<i>Message</i>	<i>Required Actions(s)</i>	<i>Points</i>
<i>Provide point value of [active target]</i>	Enter point value into Command chat box. Hit “enter.”	25
<i>Provide search radius of [active target]</i>	Click on target on moving map or target name in the Target Info. Table. Retrieve radius value from bottom right of display.	25

	Enter value into Intelligence chat box and hit “enter”	
<i>[Active target] update - set search radius to [value to the tenth decimal place, usually lower than current value]</i>	Click on target on moving map or target name in the Target Info. Table. Input new value into radius entry field at the bottom of the right screen, no unit required, and hit “enter”	25
<i>Provide latitude of [active target]</i>	Click on target on moving map or target name in the Target Info. Table. Retrieve latitude value from bottom right of display. Enter latitude, no unit required, into Intelligence chat box. Hit “enter”	25
<i>Provide longitude of [active target]</i>	Click on target on moving map or target name in the Target Info. Table. Retrieve longitude value from bottom right of display. Enter longitude, no unit required, into Intelligence chat box. Hit “enter”	25
<i>[Active target] update - set latitude and longitude to [true target latitude to thousandths decimal place] and [true target longitude to thousandths decimal place]</i>	Click on target on moving map or target name in the Target Info. Table. Input new value into latitude entry field at the bottom of the right screen, no unit required, and hit “enter” Input new value into longitude entry field, no unit required, and hit “enter.” Optionally reduce radius value to zero and hit “enter” to obtain up-to-date route times	25 (latitude) + 25 (longitude)

Note. Intel will never request information on expired targets.

Participants receive points if they correctly respond to a message within one minute, either by providing the correct value in the appropriate chat field or by correctly updating the appropriate vehicle or target parameter. Possible message types from Command and Intelligence and their associated point rewards are provided in Table A.5 and Table A.6, respectively. If a participant fails to respond to a message, or responds after a minute has elapsed, they will not receive points.

A.5.2. Sensor orientation task

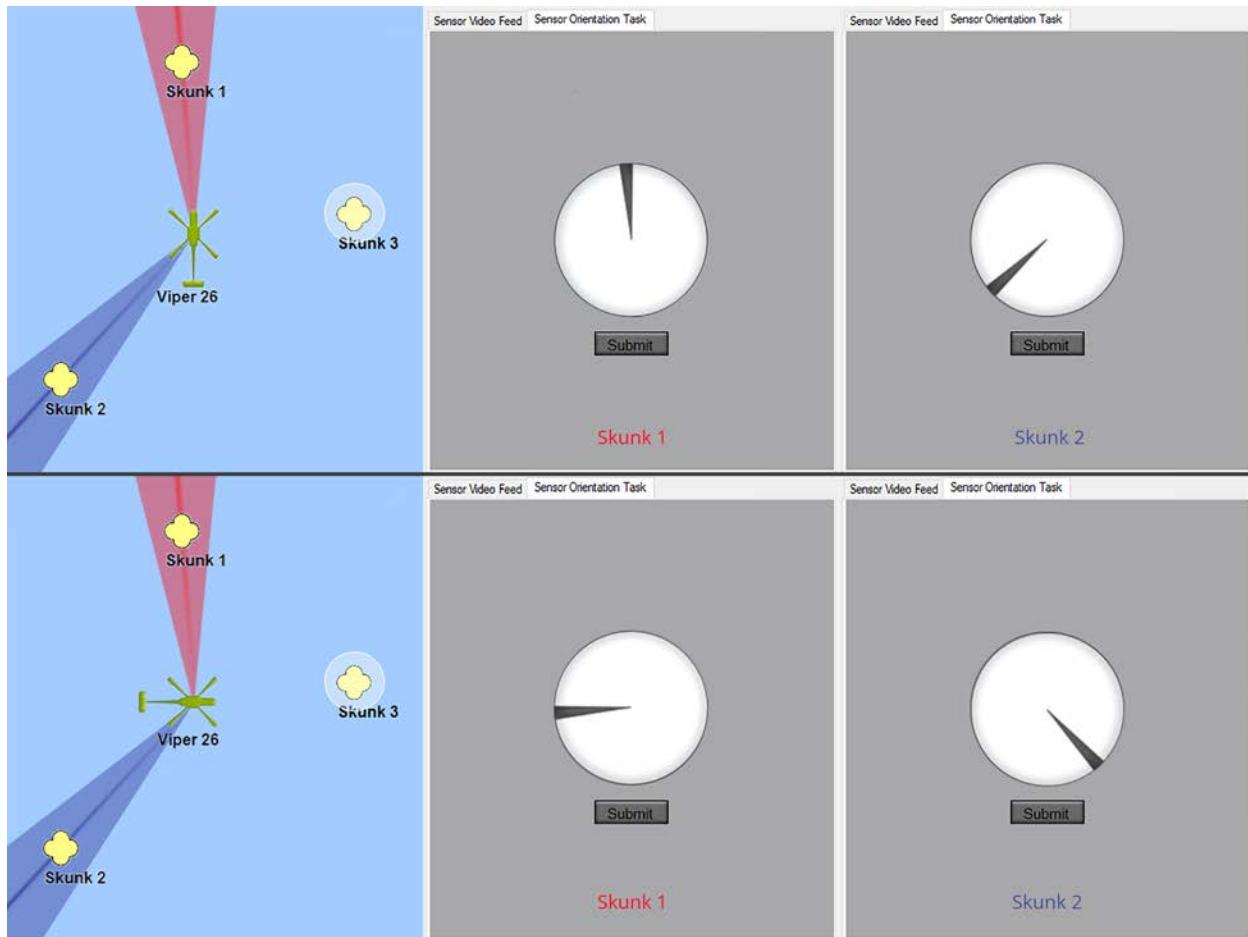


Figure A.13. Sensor orientation task.

Occasionally, Command will ask a participant to “Aim [randomized UAV]’s sensor at [nearby target].” To aim the sensor, the participant must first click on the Sensor Orientation Task tab of the specified UAV, located at the top of its Sensor Video Feed, to display its payload tool (Figure A.13). The participant may then hover over the aiming tool with their mouse, moving the mouse in a circular motion until the sensor’s directional wedge is aimed at the requested target.

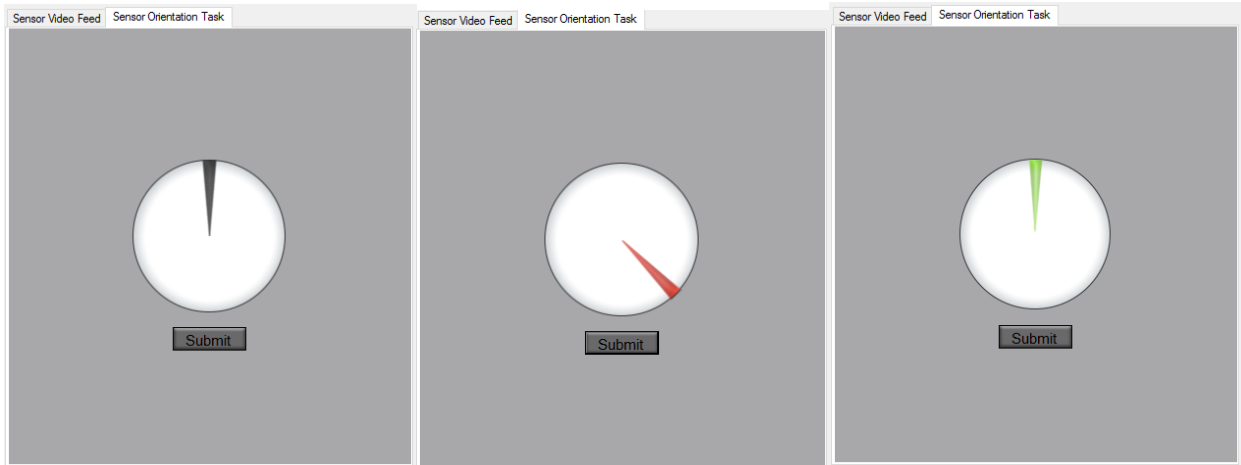


Figure A.14. Sensor orientation task operator feedback.

Once the participant is satisfied that the sensor is aimed in the correct direction, they may click on the dark gray directional wedge. A yellow outline around the wedge indicates the direction is locked in. If the participant wishes to change the direction, they can do so by clicking on the wedge again to release it. If they are happy with their current selection, they may click on the submit button to input their answer. If the wedge turns green, they selected the correct direction within 20 degrees of error and will receive 100 points. If it turns red, they did not select the correct direction and will not receive points (Figure A.14).

A.5.3. Reporting UAV position. Once a UAV is within approximately five minutes of its estimated time of arrival to its next scheduled target, as shown in the “Time of Arrival” column of the UAV Route Builder box (Figure A.12), the participant should report the position of the UAV to Command to receive additional points. In addition to the “Time of Arrival” in the Route Builder box, which is reported in mission clock time, the participant can view the countdown to target arrival for each of their UAVs by clicking on any of the UAV icons on the map. A pop-up box will appear showing “ETA” as a countdown to target arrival.

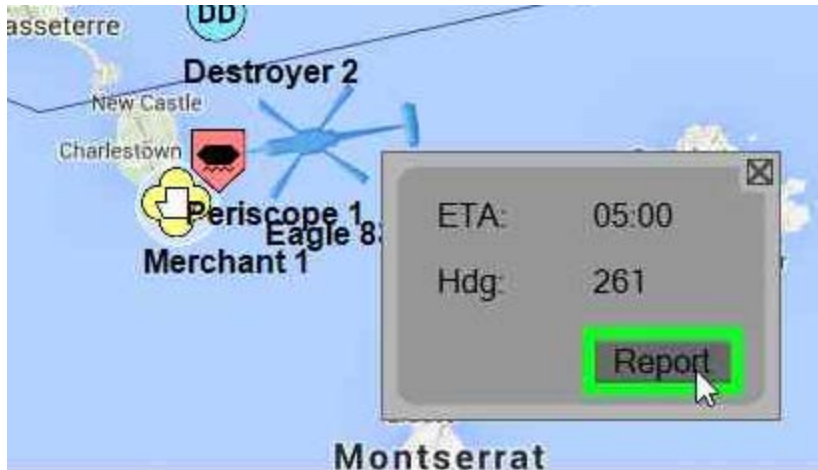


Figure A.15. UAV ETA report feature.

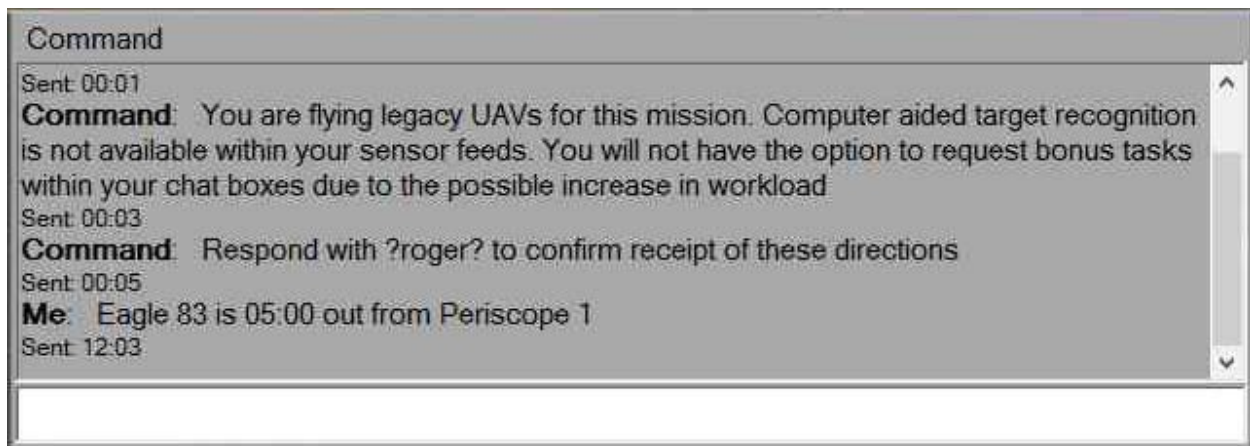


Figure A.16. UAV ETA report automatic text to Command.

To report a UAV's position, the participant must click its icon on the map and select "Report" from the drop-down box (Figure A.15). A text message with the UAV's estimated ETA will automatically be sent to Command and can be seen in the chat field (Figure A.16). If the participant reports the UAV's position within four to six minutes of its arrival, they will receive 100 points. If the participant reports their UAV's location a little too soon (up to seven minutes before its arrival) or a little too late (up to three minutes before its arrival), they will receive 50 points. They will not receive points for reporting a UAV's position at any other time.

A.5.4. Payload task. A synthetic voice alert, “started searching [target name],” will alert the participant when a UAV begins to search for a target. When the participant hears this alert, they must begin to monitor the searching UAVs simulated sensor feed.

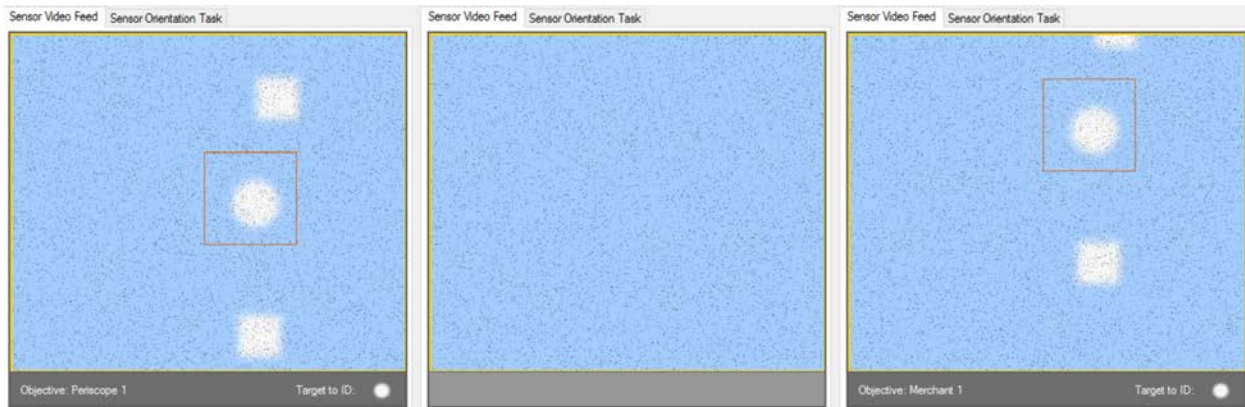


Figure A.17. Less complex payload task.

SCOUT includes automated target identification to assist participants in locating the target of interest. This tool will preselect potential targets, which are then outlined in brown (Figure A.17). The participant has until the target reaches the bottom of the sensor feed to deselect any target selections with which they do not agree. Depending on the experimental condition, this automated tool will either be quite reliable or unreliable (subject to a liberal response criterion). Although the unreliable automation variant will preselect all potential targets, it will also erroneously preselect a number of distractor targets that will result in point loss unless the participant corrects the false alarms while they are still displayed in the sensor feed. Incorrectly selected distractor targets are indicated by an auditory alert, so the potential for participant distraction and annoyance is higher in the unreliable condition.

In less complex sensor pictures, the target of interest varies across one dimension: shape. The target of interest could be a circle or a square. In Figure A.17, the participant is searching for Periscope 1, which is a circle. The participant must click on all the potential targets—all the

circles—as they appear while ignoring distractor squares. Only one circle is Periscope 1, but the participant must click on all potential targets, all the circles, to maximize their chance of locating Periscope 1.

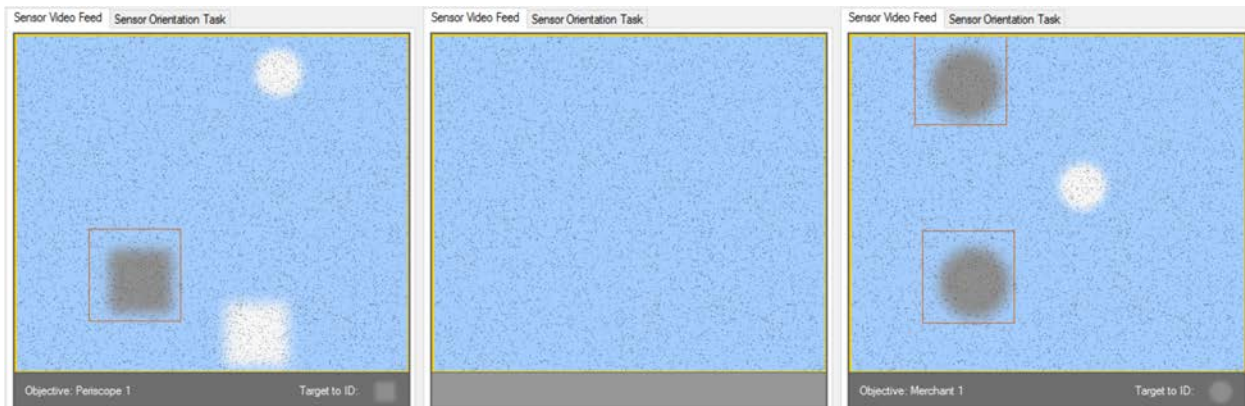
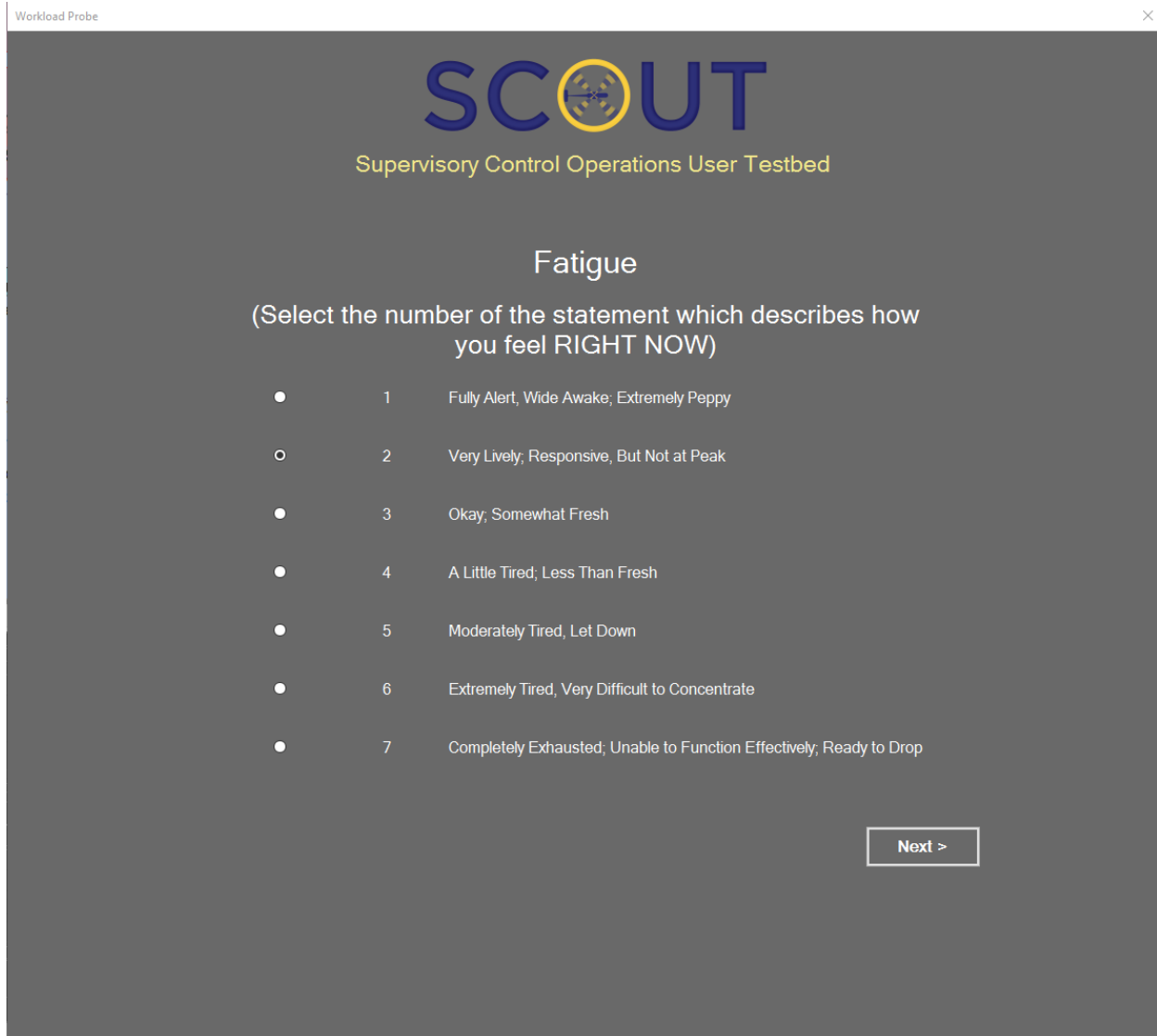


Figure A.18. More complex payload task.

In more complex sensor pictures, the target of interest varies across three dimensions: shape (circle or square), color (gray or white), and size (large or small). In Figure A.18, the participant is searching for Periscope 1, which is a small white circle. The participant must click on all potential targets, all the small white circles, to maximize their chance of finding Periscope 1.

A synthetic voice alert, “[target name] has been located,” will alert the participant once they locate the target. At this point, the sensor feed will go blank as the UAV departs the target area. If the participant misses the target, they will be given multiple additional chances to find it. If they keep missing the target, their UAV will continue to search the target search area until the target deadline is reached. At that point, the UAV will automatically break off the search and move on to its next target or waypoint if one is assigned in its Route Builder. If a subsequent target or waypoint is not assigned, it will loiter above the expired target until it is directed to another target or waypoint.

A.5.5. Fatigue and workload questionnaires.



Workload Probe ×

SCOUT
Supervisory Control Operations User Testbed

Fatigue

(Select the number of the statement which describes how you feel RIGHT NOW)

- 1 Fully Alert, Wide Awake; Extremely Peppy
- 2 Very Lively; Responsive, But Not at Peak
- 3 Okay; Somewhat Fresh
- 4 A Little Tired; Less Than Fresh
- 5 Moderately Tired, Let Down
- 6 Extremely Tired, Very Difficult to Concentrate
- 7 Completely Exhausted; Unable to Function Effectively; Ready to Drop

Figure A.19. Fatigue probe.

There will be several times throughout the mission where the mission clock will stop and a new screen, a fatigue and workload questionnaire, will appear. This quick, subjective assessment is based on the Crew Status Survey (Samn & Perelli, 1982). The first page of the pop-up screen asks the participant to rate their current level of fatigue on a seven-point scale from “fully alert” (1) to “completely exhausted” (7) (Figure A.19). The participant may make

their selection by clicking on the radio button that corresponds to their current level of fatigue. Once they finish, they may click “next” to move to the second page.

Workload Probe

SCOUT

Supervisory Control Operations User Testbed

Workload Estimate

Select the number of the statement which describes the MAXIMUM workload you experienced during the past work period in the left column and the AVERAGE workload you experienced in the second column

Maximum	Average		
<input type="radio"/>	<input type="radio"/>	1	Nothing to do; No System Demands
<input type="radio"/>	<input type="radio"/>	2	Light activity; Minimum demands
<input type="radio"/>	<input type="radio"/>	3	Moderate activity; Easily managed; Considerable spare time
<input type="radio"/>	<input type="radio"/>	4	Busy; Challenging but manageable; Adequate time available
<input type="radio"/>	<input type="radio"/>	5	Very busy; Demanding to manage; Barely enough time
<input type="radio"/>	<input type="radio"/>	6	Extremely busy; Very difficult; Non-essential tasks postponed
<input type="radio"/>	<input type="radio"/>	7	Overloaded; System unmanageable; Essential tasks undone; Unsafe

Figure A.20. Workload probe.

The second page asks participants to estimate both their average and maximum workload on a seven-point scale from “nothing to do” (1) to “unmanageable” (7) (Figure A.20). They may make their selection by clicking on the radio button that corresponds to the average amount of workload they have experienced since the last probe or the beginning of the mission, whichever came last. Once they finish, they may click “submit” to return to the mission. They may resume

the mission right away by clicking the “resume” button; otherwise, the mission will resume automatically after a 10-second pause.

A.6 Summary

Each participant will complete one walkthrough training mission, a 13:45 practice mission, and one 34:15 experimental mission with five 45-second workload/fatigue freeze probes. Before each mission block (apart from the training walkthrough), they will engage in an untimed planning period. Although the planning period is not time pressured, it takes most participants about five to 10 minutes to complete. Participants should use a combination of information from the moving map, the Target Information table, and the UAV Route Builder boxes to formulate the best plan for sending their UAVs to targets and maximizing their points.

Once the participant has decided on a plan and hit the “begin mission” button, the mission clock will start. Throughout each mission, the participant must continually replan based on new targets that show up, doing their best to monitor payload feeds, answer information requests, and make status updates with the goal of maximizing their points.

Appendix B: Low-Cost Eye Tracking

Supervisory control often involves extended periods of monitoring and, during such periods of low task load, traditional performance metrics (e.g., accuracy and response time) might not be available or even be representative of good performance. One solution for gathering a more complete picture of operator performance is to augment traditional metrics of mission performance with eye tracking metrics, such as pupil diameter. Research has shown a direct relationship between pupil diameter in millimeters and workload (Beatty, 2000). Pupil dilation in response to mental activity is a well-established phenomenon in neuropsychology. The German neurologist, Oswald Bumke, wrote in 1911:

Every active intellectual process, every psychical effort, every exertion of attention, every active mental image, regardless of content, particularly every affect just as truly produces pupil enlargement as does every sensory stimulus. (Hess, 1975, pp. 23–24).

However, it was not until 1964 that the task-evoked pupillary response was first used as a tool to investigate human cognitive processing and mental effort. Hess and Polt (1964) claimed that pupillary dilations indicate mental effort after observing a direct relationship between the difficulty of mental arithmetic problems and the magnitude of participants' pupil dilation during the problem-solving period. This relationship between the magnitude of pupil dilation and task demand or difficulty was subsequently observed in a variety of contexts: arithmetic (Bradshaw, 1968b; Payne, Perry & Harasymin, 1968); short-term memory tasks of varying load (Kahneman & Beatty, 1966); pitch discriminations of varying difficulties (Kahneman & Beatty, 1967); standard tests of "concentration" (Bradshaw, 1968a); sentence comprehension (Wright & Kahneman, 1971); paired-associate learning (Colman & Paivio, 1970; Kahneman & Peavler, 1969); imagery tasks with abstract and with concrete words (Paivio & Simpson, 1966, 1968;

Simpson & Paivio, 1968), and the emission of a freely selected motor response instead of an instructed response (Simpson & Hale, 1969) (as cited in Kahneman, 1973).

Kahneman proposed the task-evoked pupillary response as the primary measure of processing load in his effort theory of attention (Beatty, 1982; Kahneman, 1973). He justified the use of the physiological measure based on the strong empirical support for the direct relationship between pupil dilation and task demands (specifically, by citing the prior list), concluding that “the key observation that variations of physiological arousal accompany variations of effort shows that the limited capacity [to perform mental work] and the arousal system must be closely related.” (Beatty, 1982; Kahneman, 1973, p. 10). Kahneman stated that a physiological measure of mental effort must be sensitive to both between-task and within-task variations. Not only should such a measure be able to order tasks by difficulty (since more difficult tasks usually require greater mental effort), but it should also reflect transient variations in participants’ effort during task performance.

There is ample empirical support for the sensitivity of pupil diameter as a measure of momentary within-task variations in cognitive effort. Sibley, Coyne and Baldwin (2010) used pupil diameter to assess mental effort within a simulated UAV training task and found that pupil diameter decreased as performance increased.

B.1 Low-Cost Eye Tracking Systems

Historically, the high cost of eye tracking systems has limited the use of the technology in Human Factors research. High-end eye tracking systems range from \$15,000 to over \$80,000. However, in recent years, a number of low-cost eye tracking solutions have become available. These systems, which range from \$100 to \$500, are designed for use with single displays and offer a streamlined setup process.

Table B.1

Technical Specifications for Three First-Generation Low-Cost Eye Tracking Systems

	Gazepoint GP3	Eye Tribe	Tobii EyeX
Cost	\$495	\$99	\$139
Sampling Rate	60 Hz	30/60 Hz	60 Hz (estimated)
Visual Angle	0.5°–1.0°	0.5°–1.0°	–
Max. Display Size	24 in.	27 in.	27 in.
Eye Position Data	Left and Right	Left and Right	Combined
Pupil Size Data	Pixels	Pixels	None

Note. Eye Tribe defaults to a 30 Hz sampling rate and Coyne and Sibley (2016) experienced issues with duplicate packets at 60Hz.

The first generation of low cost eye trackers includes the Tobii EyeX, Gazepoint GP3, and Eye Tribe. The Gazepoint GP3 and the Eye Tribe collect data on gaze position and pupil size for both eyes.¹ The Tobii EyeX, however, only provides gaze position averaged across both eyes. Moreover, the EyeX was designed for entertainment purposes and, unfortunately, the user agreement does not permit data collection and analysis. A summary of the technical specifications for these three systems is provided in Table B.1.

The two best indicators of eye tracking data quality, and thus the quality of the eye tracking systems themselves, are the accuracy and precision of the gaze data. Accuracy, which is sometimes referred to as offset, is the difference between the true and measured gaze direction. Accuracy is measured in visual angle which is 0.5°–1.0° for both the Gazepoint GP3 and Eye Tribe systems, according to the manufacturers. Precision refers to the consistency of calculated gaze points when the true gaze direction is held constant. The precision of an eye tracker is typically measured using an artificial eye in order to estimate the magnitude of system noise or error. Gazepoint and Eye Tribe have not published precision figures, but researchers have experimentally estimated their precision. Ooms, Dupont, Lapon, & Popelka (2015) found that the

¹ Pupil data was collected in pixels here, but the updated version of the Gazepoint GP3 can report pupil diameter in millimeters.

gaze accuracy and precision of Eye Tribe was comparable to the SMI RED 250, an established, high-end system.

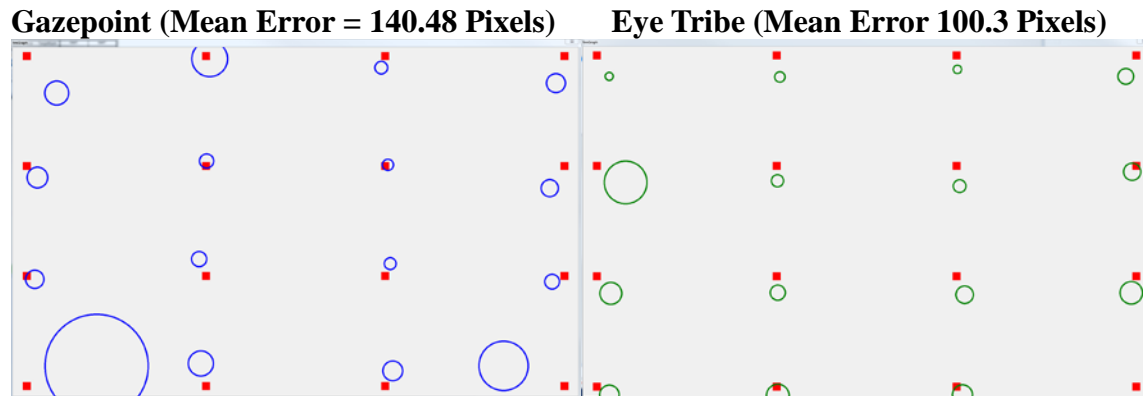


Figure B.1. Accuracy and precision for Gazepoint (left) and Eye Tribe (right) eye tracking systems on a 24-inch 1900 x 1200 display.

Similarly, Coyne, Sibley, and Sherwood (2016) found that gaze data collected using the Eye Tribe and Gazepoint GP3 systems was of sufficient accuracy and precision to be useful for Human Factors research and, on 24-inch or smaller displays, tracked gaze position almost as well as the high-cost Smart Eye Pro system. On the 24-inch 1900 x 1200 display used in the experiment, one degree of visual angle equated to approximately 49 pixels. Thus, the mean visual angle was 2.87 degrees for the Gazepoint GP3 and 2.05 degrees for the Eye Tribe (Figure B.1).

Though these experimental figures represent a higher degree of error than the values provided by the manufacturers suggest, they are inflated by lower gaze accuracy in the corners of the display. Inaccuracy in the corners is often less problematic since they are not often considered areas of interest. Moreover, the eye tracker set up represents the “worst case scenario”; placement of the eye trackers and their tripods was not controlled. Additional testing of the Gazepoint system with its optional VESA mount, focused on a smaller area of the display (the middle 90%), yielded a reduced visual angle of 1.60 degrees with minimal filtering. This

finding suggests that these systems are capable of more accurate data collection if their position is fixed center and level with respect to the monitor.

Table B.2

Low-Cost Eye Tracker Data Quality with Three Filters Applied

Filter Applied	System	Mean Error	Percent Usable Data
One Good Eye	Gazepoint GP3	140.48	94%
	Eye Tribe	100.30	79%
Two Good Eyes	Gazepoint GP3	133.58	83%
	Eye Tribe	91.67	64%
Two Good Eyes and < 200	Gazepoint GP3	95.35	69%
	Eye Tribe	82.48	60%

However, Coyne et al. (2016) found that the low-cost systems experienced more frequent data quality problems relative to the Smart Eye Pro. Depending on the filter applied, up to 40% of data were lost. Table B.2 lists the percent usable data remaining after three increasingly selective filters were applied to the raw gaze data: (1) minimal, in which packets with at least one good eye were included; (2) both eye, in which only packets with good quality reported for both left and right gaze position were included; and (3) both eye distance, which included only packets with good quality left and right gaze position within two degrees of visual angle.

Funke et al. (2016) compared the accuracy and precision of Tobii EyeX and Eye Tribe to that of three more costly eye tracking systems: Seeing Machines faceLAB, Smart Eye Pro, and Smart Eye Aurora. Like Coyne et al. (2016), they found that, while the accuracy and precision of the low-cost systems were comparable to the more expensive systems, data collection with the Tobii EyeX and Eye Tribe resulted in fewer usable gaze estimate data points due to more frequent data quality problems. They cautioned that missing data could affect estimates of the number and duration of fixations, saccadic rates, and blinks, all of which are commonly used in Human Factors research. Thus, researchers should carefully consider the relative strengths and

weaknesses of the various systems and their suitability for their specific research effort (Funke et al., 2012; Holmqvist, Nystrom, & Mulvey, 2012).

One notable weakness of low-cost eye tracking systems is their relatively low sampling frequency (30–60 Hz) relative to high-cost systems (often 250 Hz and above), which limits their suitability for research where rapid eye movements are of interest (Ooms et al., 2015). For example, experimenters who conduct reading research often track saccades; saccades are the ballistic, conjugate eye movements that occur between fixations, such as those that characterize eye movement while reading. The speed of these eye movements necessitates the use of a more expensive, high-speed system with sampling rates of 500 Hz or more to get meaningful data (Poole & Ball, 2006; Rayner & Pollatsek, 1989). Fortunately for Human Factors researchers, the 60 Hz sampling rate of many low-cost eye trackers is sufficient for human-computer interaction and usability studies, including those which would typically occur in UAV supervisory control test beds.

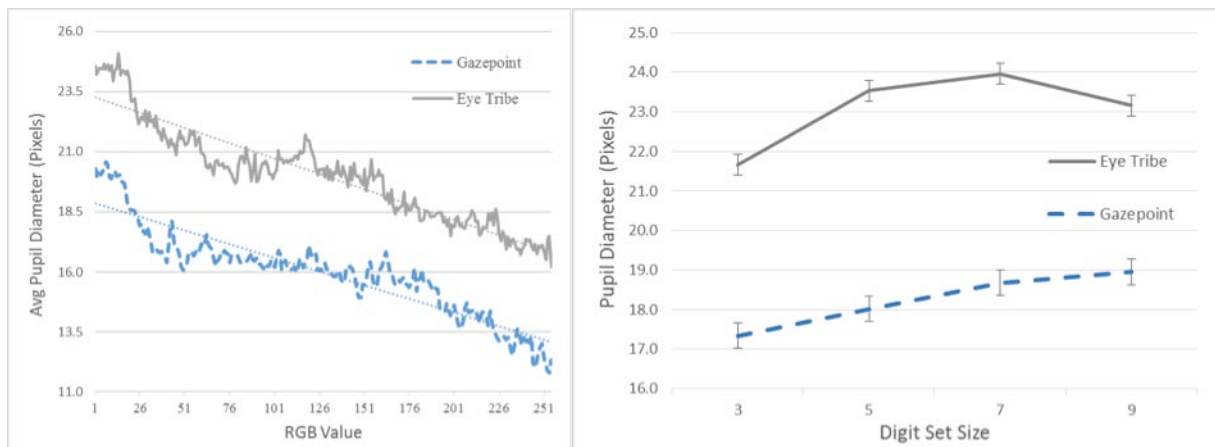


Figure B.2. Pupillary response to increasing screen luminance (left) and workload (right) as measured by the Gazeport GP3 and Eye Tribe systems. *Note.* Reprinted from “Investigating the use of Two Low Cost Eye Tracking Systems for Detecting Pupillary Response to Changes in Mental Workload,” by Coyne and Sibley (2016). Reprinted with permission.

While a number of studies have investigated the accuracy and precision of gaze data collected using low-cost eye trackers, less is known about the ability of these devices to collect non-gaze behaviors, such as pupil size. Coyne and Sibley (2016) found the Eye Tribe and Gazepoint GP3 systems sufficiently sensitive to capture changes in pupil size in response to both screen luminance and mental effort on a digit span task (Figure B.2).

However, unlike the Smart Eye Pro and similar high-cost systems that measure pupil size in millimeters, both low-cost systems output pupil size in pixels. The pixel-counting method of measurement is potentially problematic because a participant's observed pupil size, the number of pixels their pupils occupy in the camera image, can be confounded by their gaze angle and head position. However, this problem can be mitigated by software that uses an ellipse-fitting method to measure pupil size; the ellipse-fitting method defines pupil size as the length of the major axis of an ellipse fitted to the pupil image and is thus not affected by perspective distortion (Klinger, 2008; Wang, 2011). Both Eye Tribe and Gazepoint have recently released low-cost systems capable of measuring pupil size in millimeters.

Overall, while low-cost eye trackers are not quite as accurate and experience more frequent data quality problems relative to high-end systems, research suggests that these devices may be able to provide meaningful data in applied settings.

