

# The Design of a HSMM-Based Operator State Monitoring Display

by

Ryan W. Castonia

S.B. Aerospace Engineering  
Massachusetts Institute of Technology, 2010

Submitted to the Department of Aeronautics and Astronautics  
in partial fulfillment of the requirements for the degree of

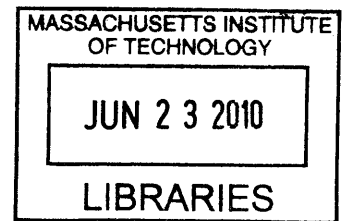
Master of Science in Aeronautics and Astronautics

at the

Massachusetts Institute of Technology

June 2010

**ARCHIVES**



©2010 Massachusetts Institute of Technology. All rights reserved.

Signature of Author: \_\_\_\_\_

*RW*  
Ryan W. Castonia  
Department of Aeronautics and Astronautics  
May 21, 2010

Certified by: \_\_\_\_\_

*MLC*  
M. L. Cummings  
Associate Professor of Aeronautics and Astronautics  
Thesis Supervisor

Accepted by: \_\_\_\_\_

*EH*  
Eytan H. Modiano  
Associate Professor of Aeronautics and Astronautics  
Chair, Committee on Graduate Studies



# **The Design of a HSMM-based Operator State Monitoring Display**

by

Ryan William Castonia

Submitted to the Department of Aeronautics and Astronautics on May 21, 2010 in partial fulfillment of the requirements for the degree of Master of Science in Aeronautics and Astronautics

## **Abstract**

This thesis presents the development of and the findings from the design and evaluation of a hidden semi-Markov model (HSMM)-based operator state monitoring display. This operator state monitoring display is designed to function as a decision support tool (DST) for the supervisor of a small team of operators (between 2 and 4 team members) that are each monitoring and controlling multiple highly autonomous heterogeneous unmanned vehicles (UVs). Such displays are important in real-time, mission-critical complex systems because both operator and vehicle state monitoring are difficult, and failure to appropriately handle emerging situations could have life or mission-critical consequences.

Recent research has shown that HSMM-based models can be used to model the time-sensitive behavior of operators controlling multiple heterogeneous UVs. Because this method requires substantial knowledge in probability theory to understand, the difficulty lies in the accurate, useful display of the HSMM information to a team commander in the field. It must be assumed that the team commander generally does not have the required background in stochastic processes to understand the method and may be biased in interpreting probabilistic functions. This further increases the difficulty of the proposed method. In this thesis, a cognitive task analysis (CTA) was performed to determine the functional and information requirements of the DST, and a human systems engineering design process was used to develop a prototype display. A human subject experiment was then conducted to test the effectiveness of the DST.

The DST was shown to improve team supervisor performance in terms of increased decision accuracy, decreased incorrect interventions, and decreased response times in single alert scenarios. The DST was also shown to decrease the number of incorrect interventions, while having no affect on decision accuracy and total response time scenarios when the supervisor faced multiple simultaneous alerts. Additionally, the DST was not shown to increase operator mental workload, as measured by a secondary task, for any of the scenarios. Overall, the results suggest that HSMM-based operator state modeling can be effectively utilized in a real-time DST for team supervisors.

While this research was focused on a team supervisor of multiple operators each independently controlling multiple heterogeneous UVs, the results are generalizable, and any research in time-critical team HSC domains may benefit from this work.

Thesis Supervisor: M. L. Cummings

Title: Associate Professor of Aeronautics and Astronautics



## Acknowledgements

First and foremost, I must express my deepest gratitude, love, and awe that God has continuously blessed me beyond my wildest dreams. I pray that I continuously become more obedient as I seek to more fully demonstrate His love through my thoughts, words, and actions.

Thank you, Professor Cummings, for your willingness to work with me through every situation that I brought your way. Through all of the varied Air Force requirements, my leaving for the summer to get married, and several other commitments that kept me from always being in the lab like a good grad student, you always understood and supported me. I am proud to have worked for you and look forward to keeping in touch.

Yves, working closely with you over the past couple years has been fantastic. You were always there to brainstorm with me and provide substantial, effective feedback every step of the way (except for when you were out of the country teaching SCUBA or attending weddings in castles...) I'm glad we were able to share so many experiences outside of the lab together such as Reach the Beach, WODs, and volleyball. There is no way that I could have done this without you, and I look forward to hearing about you becoming Dr. Boussemart! Thank you!

To all of my fraternity brothers and close friends at MIT, you have made my last five years an amazing experience. You are what make MIT the finest institution in the world. Thank you for your support and for so many memories.

To Hank, Jonathan, and all the UROPs that worked on this project along the way, thanks for being part of the team! I hope you enjoyed your work and hope you can apply what you learned to your future projects.

To the Great 8 and the rest of the HALiens, congratulations and good luck! Working in HAL has been an amazing adventure. I will always reminisce of volleyball in front of building 33, spirit dances at lab meetings, countless reminders about how bad we all are at writing, and the 2010 US Olympic dominance (except for that last hockey game...)

To my mother and father, I thank you for your unwavering support for the past 23+ years. Thank you for encouraging me to always aim for the top. You have always made sure that I knew I could do anything I put my mind to as long as I committed, worked hard, and never gave up. I am the man I am today because of you. I thank you more than I can express in mere words. I love you, and look forward to our adventures in trying to catch up with each other around the world!

Rachel, I have missed you deeply during my college and your high school years. I was so happy that you were able to visit us during your spring break in 2010! You know that you will always have a place to stay wherever we are, so visit often! Love ya!

Tirzah, my beautiful, caring, charming, loving, amazing wife, words cannot express the love and respect I have for you. Together, we have tackled one "impossible" task after another without wavering. We might not always take the easy road, but we take the road that is right for us. I cherish every moment that I get to spend with you. I love you z-hundred!

Thanks to Boeing Phantom Works for sponsoring this research.



# Table of Contents

Abstract.....	3
Acknowledgements.....	5
Table of Contents.....	7
List of Figures.....	9
List of Tables.....	11
List of Acronyms.....	13
1. Introduction.....	15
1.1 Problem Statement.....	17
1.2 Research Objectives.....	17
1.3 Thesis Organization.....	18
2. Background.....	19
2.1 Team Monitoring.....	19
2.2 Human Behavior Pattern Recognition and Prediction.....	22
2.2.1 Statistical Learning Techniques.....	23
2.2.2 Bayesian Methods.....	23
2.2.3 HSMM-based Operator Modeling.....	24
2.3 Display of Probabilistic Data.....	26
2.4 Background Summary.....	27
3. Design of a HSMM-based Operator State Decision Support Tool.....	29
3.1 Cognitive Task Analysis.....	29
3.1.1 Decision Ladders.....	30
3.1.2 Display Requirements Checklist.....	33
3.1.3 Cognitive Task Analysis Summary.....	34
3.2 HSMM-based Operator State Decision Support Tool.....	34
3.2.1 Interaction Frequency Plot.....	35
3.2.2 Model Accuracy Prediction Plot.....	37
3.2.3 Confidence History Plot.....	40
3.2.4 Status Panel.....	43
3.3 Summary.....	47
4. Experimental Evaluation.....	49
4.1 Participants.....	49
4.2 Apparatus.....	49
4.2.1 Testing Environment Layout.....	50
4.2.2 Operator Interface: RESCHU.....	51
4.2.3 Team Supervisor Interface: HSMM-based Operator State Decision Support Tool.....	53
4.2.4 Secondary Task Display.....	54
4.3 Experimental Design.....	55
4.3.1 Independent Variables.....	56
4.3.2 Dependent Variables.....	57
4.4 Procedure.....	58
4.5 Experimental Evaluation Summary.....	59
5. Results and Discussion.....	61

5.1	Decision Accuracy .....	61
5.2	Incorrect Interventions .....	63
5.3	Response Time.....	64
5.4	Secondary Task Ratio .....	67
5.5	Subjective DST Understanding.....	68
5.6	Discussion of Experimental Findings .....	70
5.6.1	Decision Accuracy .....	70
5.6.2	Incorrect Interventions .....	71
5.6.3	Response Time.....	72
5.6.4	Secondary Task Ratio .....	73
5.6.5	Subjective Feedback .....	74
5.6.6	Summary of Results.....	75
6.	Conclusions.....	77
6.1	Supervisor Performance .....	77
6.2	Display of Probabilistic Data .....	78
6.3	Future Work.....	78
6.4	Thesis Summary.....	79
	References.....	81
	Appendix A: Decision Ladders.....	87
	Appendix B: Decision Ladders with Display Requirements for Decision Support Tool .....	89
	Appendix C: Scenario Descriptions and Test Matrix .....	91
	Appendix D: Human Subject Post Experiment Questionnaire .....	93
	Appendix E: Human Subject Consent Form.....	95
	Appendix F: Human Subject Demographic Survey .....	99
	Appendix G: Training Slides – DST User .....	103
	Appendix H: Training Slides – Non DST User .....	111
	Appendix I: Supporting Statistics .....	115



# List of Figures

Figure 1.1: US Navy Dahlgren Integrated Command Environment Lab  
(<http://www.navsea.navy.mil/nswc/dahlgren/default.aspx>) ..... 15

Figure 2.1: Literature review visualization..... 19

Figure 2.2: Yerkes-Dodson Law adapted to performance (Hancock & Warm, 1989)..... 21

Figure 2.3: Large-screen mission status display (Scott, et al., 2007; Scott, et al., 2009)..... 22

Figure 2.4: A three-state hidden semi-Markov model (Castonia, 2010) ..... 25

Figure 3.1: Systems engineering process (Blanchard & Fabrycky, 1998) ..... 29

Figure 3.2: A decision ladder with human behavior heirachies shown (Rasmussen, 1983) ..... 30

Figure 3.3: Display requirements for the “Is There a Problem?” decision ladder ..... 33

Figure 3.4: DST (Castonia, 2009) ..... 35

Figure 3.5: Interaction Frequency plot ..... 35

Figure 3.6: Interaction Frequency plot with icons shown ..... 36

Figure 3.7: Interaction frequency plot with icons and grid shown ..... 36

Figure 3.8: Model Accuracy Prediction plot ..... 37

Figure 3.9: Discrete variation of the Model Accuracy Prediction display (Castonia, 2009)..... 39

Figure 3.10: Linear variation of the Model Accuracy Prediction display (Castonia, 2009)..... 39

Figure 3.11: Curved variations of the Model Accuracy Prediction display (Castonia, 2009)..... 40

Figure 3.12: Confidence History plot ..... 40

Figure 3.13: Model Accuracy Prediction plot with applicable Confidence History plot values ..... 41

Figure 3.14: Annotated Confidence History plot ..... 42

Figure 3.15: Confidence History plot with 3 min prediction not shown ..... 43

Figure 3.16: Status Panel ..... 43

Figure 3.17: DST with UIN's highlighted ..... 44

Figure 3.18: Alert panel..... 46

Figure 3.19: Status Panel with alert shown ..... 46

Figure 3.20: Annotated DST with Design Requirements (DR) shown (Castonia, 2009)..... 47

Figure 4.1: Testing environment layout ..... 50

Figure 4.2: Operator workstation (RESCHU and DST)..... 51

Figure 4.3: Annotated screenshot of the operator display – RESCHU (Nehme, 2009) ..... 52

Figure 4.4: RESCHU – payload camera view (visual search task) ..... 53

Figure 4.5: DST screenshot with alert shown..... 54

Figure 4.6: Large-screen wall display with scenario 1 video shown..... 55

Figure 5.1: Mean Decision Accuracy per subject.....	62
Figure 5.2: Mean Number of Incorrect Interventions per subject .....	64
Figure 5.3: Mean first Response Times per subject (individual scenarios).....	65
Figure 5.4: Mean first Response Times per subject (grouped scenarios).....	65
Figure 5.5: Mean second Response Times (total Response Time) per subject .....	66
Figure 5.6: Mean response interval per subject.....	66
Figure 5.7: DST Understanding - Likert scale box plots.....	69
Figure A.1: “Is there a problem?” Decision Ladder .....	87
Figure A.2: “What should I do to solve the problem?” Decision Ladder.....	88
Figure B.1: “Is there a problem?” Decision Ladder with display requirements.....	89
Figure B.2: “What should I do to solve the problem?” Decision Ladder with display requirements .....	90

## List of Tables

Table 3.1: Display requirements for DSTs that provide predictions of future operator performance .....	34
Table 3.2: UIN Descriptions .....	45
Table 4.1: Secondary Task target utterance and occurrences .....	55
Table 4.2: Experimental scenario descriptions .....	58
Table 5.1: Decision Accuracy results .....	62
Table 5.2: Incorrect Interventions results .....	63
Table 5.3: Early Interventions.....	67
Table 5.4: Secondary Task Ratio results.....	67
Table 5.5: DST Understanding - Likert scale descriptive statistics .....	68



## List of Acronyms

ATC	Air Traffic Control
CTA	Cognitive Task Analysis
DR	Design Requirement
DoD	Department of Defense
DST	Decision Support Tool
HALab	Humans and Automation Lab
HALE	High Altitude Long Endurance Unmanned Vehicle
HMM	Hidden Markov Model
HSC	Human Supervisory Control
HSMM	Hidden Semi-Markov Model
ISR	Intelligence, Surveillance, and Reconnaissance
MAS	Model Accuracy Score
MIT	Massachusetts Institute of Technology
POMDP	Partially Observable Markov Decision Process
RESCHU	Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles
SA	Situational Awareness
UIN	User-Initiated Notification
UAV	Unmanned Aerial Vehicle
UUV	Unmanned Underwater Vehicle
UV	Unmanned Vehicle



## 1. Introduction

The movement toward single operators controlling multiple unmanned vehicles (UVs) is gaining momentum throughout academia and industry as a whole (Cummings, et al., 2007; DoD, 2007b, 2009). As this transition occurs, multiple operators will likely work together, under the leadership of a team supervisor, to carry out coordinated tasks similar to present-day air traffic control (ATC) settings. An example of a similar human supervisory control (HSC) environment in which multiple operators are working together is the US Navy's Integrated Command Environment Lab, shown in Figure 1.1.



**Figure 1.1: US Navy Dahlgren Integrated Command Environment Lab**  
(<http://www.navsea.navy.mil/nswc/dahlgren/default.aspx>)

There are many ways in which multiple operators and a supervisor could interact in these futuristic scenarios, one of which places the supervisor in a HSC role over the operators. This creates a complex HSC system in which decision support tools (DSTs) have been shown to improve supervisor performance (Mitchell, Cummings, & Sheridan, 2004). This thesis focuses specifically on developing a decision support tool for a supervisor of multiple, independent operators, i.e. operators are assigned their own areas of responsibility which do not significantly overlap. ATC sector control is an example of this type of relationship.

While there has been substantial effort in designing and developing displays for the individual operators controlling multiple UVs, there has been little research into what type of displays a team supervisor requires to most efficiently supervise the team (Scott, et al., 2007), even though it is recognized that the role of the team commander is critical (Burns, 1978; Hackman, 2002). Within the development process of these types of futuristic displays, it has become evident that the common tasks of navigation, monitoring the health and status of the UVs, and operating multiple sensors provide each operator with vast amounts of information (Nehme, et al., 2006). It is unrealistic to provide the supervisor with all this information from each operator without expecting the supervisor to become overloaded. Therefore, automation may be useful in assisting a team supervisor in the assessment of operator performance. Of the little research that begins to address the issue of supervisor displays, supervisors in human operator-unmanned vehicle teaming environments are only being provided with situational awareness tools capable of identifying *current* problems. However, supervisory decision support tools that attempt to predict the onset of possible *future* problems have not been explored (Scott, Sasangohar, & Cummings, 2009). This is another area where automation may be helpful, particularly in real-time complex systems, such as UV command and control systems, where both operator and vehicle state monitoring are difficult, and failure to appropriately handle emerging situations could have catastrophic consequences (Leveson, 1986).

Accurately detecting current and predicting possible future problems requires a thorough understanding of what “normal” operator behavior entails. This is a difficult task in a time-pressured command and control environment that contains significant uncertainty and unanticipated events. Recent research has shown that hidden semi-Markov models (HSMMs) can be used to model the time-sensitive behavior of operators controlling multiple heterogeneous UVs (Huang, 2009). Such models can both describe and predict operator behaviors. The prediction, expressed as a sequence of the most likely future operator behaviors, can be compared to learned patterns of normal operator behavior. The supervisor can then be alerted when the predicted behavior deviates from, or is expected to deviate from, the norm. It is important to note however, that such models cannot detect whether an abnormal pattern is good or bad, only that it is different, which is why the supervisor is needed to make the important decision of whether the abnormal behavior is detrimental and keep the team operating efficiently. Thus, full automation is not feasible; the human must remain part of the system.

Additionally, HSMM modeling methods require substantial knowledge in probability theory to both implement and understand. Thus, the difficulty lies in the accurate, useful display of the information to a team commander in the field who generally does not have the required background in stochastic processes and may be biased in interpreting probabilistic functions (Tversky & Kahneman, 1974).



This thesis focuses on the creation and evaluation of such a display, specifically utilizing the probabilistic output of the aforementioned HSMM-based operator model, in order to support a team commander with little to no background in statistical inferential reasoning. Design requirements were derived from a cognitive task analysis (Crandall, Klein, & Hoffman, 2006), the DST was developed, and a human subject experiment was conducted to test the effectiveness of the DST. Results show the DST improved team supervisor performance in terms of increased Decision Accuracy, decreased Incorrect Interventions, and decreased Response Times in single alert scenarios. In scenarios where the supervisor faced multiple, simultaneous alerts, the DST was shown to decrease the number of Incorrect Interventions and have no affect on Decision Accuracy or total Response Time. Overall, the results validate that HSMM-based operator state modeling can be effectively utilized in a real-time DST for team supervisors. While this research was focused on a team supervisor of multiple operators each independently controlling multiple heterogeneous UVs, the results are generalizable and any research in time-critical team HSC domains may benefit from this thesis.

### ***1.1 Problem Statement***

In order to leverage the recent advances in HSMM-based operator modeling techniques to improve team performance, it is first necessary to determine the best method for employing these techniques. Such models could operate either online as a real-time DST or offline as a post hoc evaluation tool. This thesis focuses on the development and evaluation of a real-time DST to determine whether team supervisors with a HSMM-based DST perform better than supervisors without the DST in terms of correctly identifying current and possible future problems, solving current problems efficiently and accurately, correctly preventing future problems from occurring, and limiting unnecessary interventions. This thesis also attempts to address the question of how to best display complex probabilistic data, specifically the HSMM-based operator model output, so that someone with little to no background in statistical inferential reasoning may efficiently and accurately make time-critical decisions.

### ***1.2 Research Objectives***

In order to address the problem statement, the following research objectives were posed:

- **Objective 1: Conduct a cognitive task analysis for a supervisor of multiple operators who are each independently controlling a set of highly autonomous UVs.** A cognitive task analysis (CTA) was conducted in order to understand the decisions the supervisor would need to make in this domain and determine what information the supervisor would need in order to quickly and

accurately make those decisions. Design requirements were then derived from the CTA in order to focus the design cycle. The CTA process and results are highlighted in Chapter 3.

- **Objective 2: Develop a HSMM-based operator state decision support tool.** The design requirements derived from the CTA were used to design a HSMM-based operator state decision support tool, herein referred to as the DST. The DST takes into account the applicable literature in Chapter 2 and a discussion of the design of the DST is included in Chapter 3.
- **Objective 3: Evaluate the effectiveness of the HSMM-based operator state decision support tool in a simulated team supervisory scenario.** In order to achieve this objective, a human subject experiment was conducted in which subject performance while using the DST was compared to subject performance in a standard team HSC setting. Details about the design of this experiment are included in Chapter 4. The experimental results and discussion are included in Chapter 5.

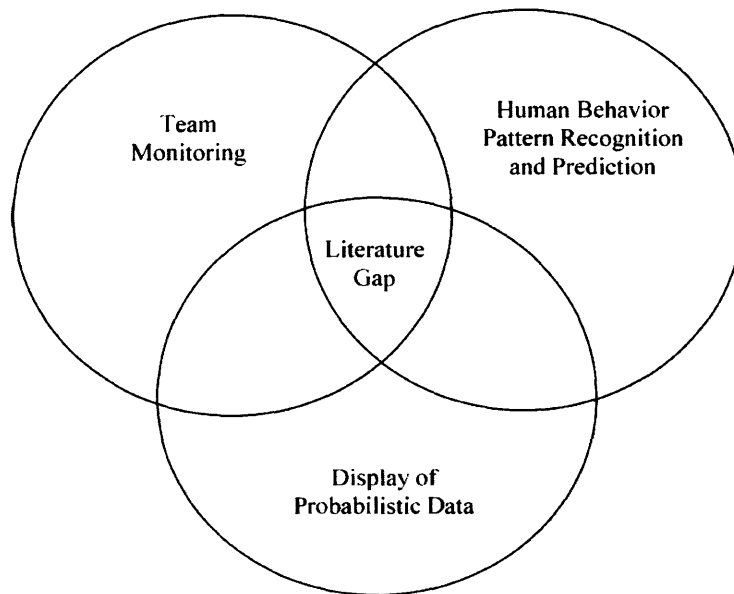
### ***1.3 Thesis Organization***

This thesis is organized into the following six chapters:

- Chapter 1, *Introduction*, provides the motivation for this research, problem statement, and research objectives of this thesis.
- Chapter 2, *Background*, presents the findings of a literature review focused on team monitoring, human behavior pattern recognition and prediction, and the display of probabilistic data. This chapter identifies the current research gap that this thesis addresses.
- Chapter 3, *Design of a HSMM-based Operator State Decision Support Tool*, explains the method used for developing the DST. The CTA is described, as well as how design requirements (DR) were derived from the process. This chapter then highlights the final design of the DST and describes how the interface meets the design requirements that were derived from the CTA.
- Chapter 4, *Experimental Evaluation*, presents the human subject experiment used to test the effectiveness of the DST in a simulated team supervisory control scenario. Descriptions of the participants, apparatus, experimental design, and procedure are included.
- Chapter 5, *Results and Discussion*, explains the results of the experimental evaluation including the dependent variables of Decision Accuracy, Incorrect Interventions, Response Time, Secondary Task Ratio, and subjective feedback.
- Chapter 6, *Conclusion*, reviews the answers to the research questions, discusses the contribution of this work, suggests design considerations and changes for future iterations of team supervisor DSTs, and identifies areas for future research.

## 2. Background

This thesis provides a team supervisor of multiple, independent operators with a DST in order to best maintain and improve team performance in a HSC setting. A significant body of research exists on team performance assessment in supervisory control settings, including team monitoring, but no clear consensus has emerged regarding how to do this effectively. Work has also been produced on the topic of human behavior pattern recognition and prediction, as well as human difficulties in understanding probabilistic data. Unfortunately, little literature exists on the topic of providing a team supervisor with real-time team monitoring support in general. No known research discusses the use of probabilistic models to provide team monitoring support, although recent work has focused on providing an individual operator with probabilistic-based decision support in port and littoral zone surveillance (Rhodes, et al., 2009). These areas of research are discussed in more detail below, and the void that exists at the intersection, providing a team supervisor with a DST that utilizes probabilistic operator modeling, is emphasized (Figure 2.1).



**Figure 2.1: Literature review visualization**

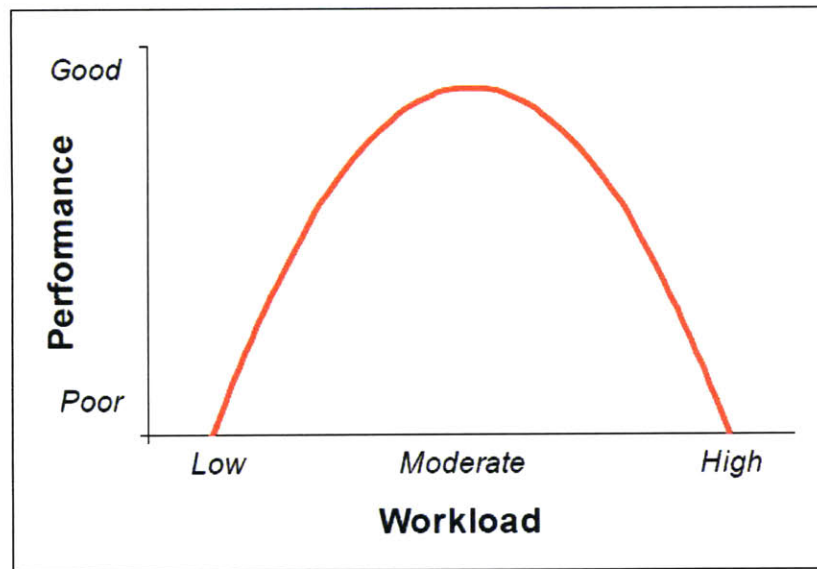
### ***2.1 Team Monitoring***

For this thesis, a team is defined as “a distinguishable set of two or more people who interact dynamically, interdependently, and adaptively toward a common and valued goal/object/mission, who have each been assigned specific roles or functions to perform, and who have a limited lifespan of

membership” (Salas, et al., 1992, p. 4). The concept of team members having a common valued goal is critical as it results in team members working together for the completion of that goal and not just out of self-interest. This concept of a common goal leads to team members assisting each other and improving overall team performance as they build a “shared mental model” of the current situation, goal, and team performance (Cannon-Bowers, Salas, & Converse, 1993).

Team monitoring, both from group members monitoring each other and from a team supervisor monitoring the group, has also been shown to improve team performance by helping a group integrate related task activities, identify appropriate interruption opportunities, and notice when a team member requires assistance (Gutwin & Greenberg, 2004; Pinelle, Gutwin, & Greenberg, 2003). While previous literature has indicated that having a constant team monitoring system in place may negatively influence operator behavior (Brewer & Ridgway, 1998; Guerin & Innes, 1993), many fields employ constant monitoring systems, such as ATC, aircraft flight operations, and generally all mission and life-critical HSC settings. Since UV crews and many military operation teams already operate under monitoring systems (DoD, 2007b), it is assumed that the team in the multiple UV scenario posed by this work is not subject to adverse social facilitation influences from a constant monitoring system. Instead, team supervisors must focus on the positive influence they can have on team performance through direct action, as well the establishment of team culture, trust, and dedication. Specifically, for a supervisor monitoring a team, determining the correct timing and number of times to intervene in team behaviors are the key factors to optimizing team performance (Brewer & Ridgway, 1998; Irving, Higgins, & Safayeni, 1986; Mitchell, et al., 2004).

This key decision, deciding when to intervene, is nontrivial. Operator performance is often difficult to infer from only observing physical actions and interactions with a computer interface, which may even be hidden from the team supervisor in HSC settings. The task becomes more difficult when the supervisor must merge information from a variety of sources, often in a time sensitive environment, to evaluate team performance. Due to such a high mental workload, the supervisor is subject to poor performance from operating on the far right side of the Yerkes-Dodson curve shown in Figure 2.2 and may make costly incorrect decisions (Hancock & Warm, 1989; Yerkes & Dodson, 1908). It has been shown that performance significantly degrades when supervisors are tasked beyond 70% utilization, or busy time (Cummings & Guerlain, 2007; Rouse, 1983; Schmidt, 1978). Therefore, the key decision of deciding when to intervene can benefit greatly from the addition of automation to offload some of the mental workload in high workload settings.



**Figure 2.2: Yerkes-Dodson Law adapted to performance (Hancock & Warm, 1989)**

One way automation can assist the supervisor is through real-time monitoring that provides a better understanding of operator and team performance through the use of a decision support tool, seen as critical in most HSC settings (Mitchell, et al., 2004). Since researchers recognize that the role of the team supervisor can be critical in improving team performance (Burns, 1978; Hackman, 2002), the advancement of supervisor decision support displays may also play a critical role in improving team performance. This advancement carries great significance in real-time complex systems where state monitoring is difficult and incorrect decisions could have catastrophic consequences (Leveson, 1986).

Recent studies that have researched team supervisor interfaces have focused on large-screen situation and activity awareness displays and the design of an Interruption Recovery Assistance display (Scott, et al., 2007; Scott, et al., 2009; Scott, et al., 2008). The Interruption Recovery Assistance display was not shown to have a significant impact on supervisor performance (Scott, et al., 2008), while the large-screen situation and activity awareness displays were shown to have high usability and effectiveness ratings through an exploratory study (Scott, et al., 2009). Of particular interest is the operator performance panel outlined by the red dashed box in Figure 2.3. This continuous operator performance monitoring display was part of a large screen mission status display and served as the original motivation for this thesis. It was used by experimental subjects to aid in the supervision of three simulated operators, each controlling three UVs in order to provide surveillance of a specific area of interest. The operator performance display was based on simple system metrics, such as the number of expected targets in each threat area. Results showed this operator performance panel was useful, and it was hypothesized that team performance could

benefit from the integration of human behavior pattern recognition and prediction techniques in order to improve this display. Different types of human behavior pattern recognition and prediction techniques, discussed in the next section, were then analyzed in the context of creating a real-time continuous monitoring DST that could improve team performance in HSC settings.

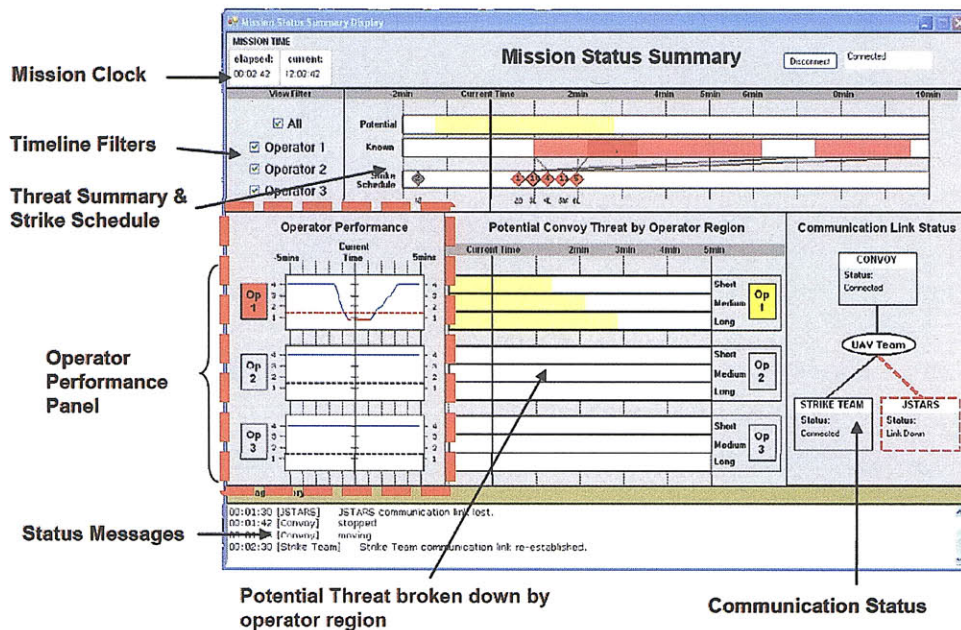


Figure 2.3: Large-screen mission status display (Scott, et al., 2007; Scott, et al., 2009)

## 2.2 Human Behavior Pattern Recognition and Prediction

While some work has focused on using pattern recognition techniques for human behaviors such as intrusion detection (Terran, 1999) and navigation (Gardenier, 1981), relatively little work has been done within the HSC realm. Within this domain, human actions have been shown to be appropriately modeled by serial processes because humans can solve only one complex problem at a time (Broadbent, 1958; Nehme, 2009; Welford, 1952). Therefore, such human behavior can be analyzed with time series analysis techniques. This technique departs from previous qualitative methodologies that focus on descriptive models and have no predictive capability (Hackman, 2002; Klein, 1999). The time series analysis techniques are beneficial because they provide a formal, quantitative basis for describing human behavior patterns and for predicting future actions. These techniques are primarily categorized as either statistical learning or Bayesian techniques and will now be discussed.

### ***2.2.1 Statistical Learning Techniques***

The simplest time series analysis technique belonging to the statistical learning category relies on descriptive statistics of operator behavior (Olson, Herbsleb, & Rueter, 1994). An example is reporting overall mean performance scores for UAV operators without showing how the performance scores changed throughout the mission (Nehme, et al., 2008). Descriptive statistics are useful in the analysis of high level mission performance but do not discriminate the temporal evolution throughout the mission. Other statistical learning techniques are based on complex neural networks that require supervised training (Polvichai, et al., 2006), which is a learning technique where a human has to parse and label operator actions into a priori network outputs. These neural networks tend to have good identification power. Unfortunately, supervised training requires a large amount of prior knowledge about the system as a whole in order to assign a specific network output to a specific behavioral input. An example is the automated written character recognition system the US Postal Service uses to identify hand-written zip codes (LeCun, et al., 1989).

A technique that does not require supervised learning is algorithmic data mining, and it has been used in many domains, such as healthcare, marketing, and fraud detection (Witten & Frank, 2005). The goal of data mining is to find correlation between variables (Kay, et al., 2006). This method is able to parse out common trends from large amounts of data but is computationally intensive and at risk of finding correlation between unrelated parameters. Similar to algorithmic data mining, exploratory sequential data analysis (Sanderson & Fisher, 1994) utilizes data-mining-like techniques with a focus on time series. This approach results in a less computationally intensive method that shares the same major drawback as data mining; there is a non-trivial risk of finding correlation between unrelated parameters. An example of sequential data analysis is an attempt to identify the patterns of behavior in medical students that cause failure in determining a genetic pedigree (Judd & Kennedy, 2004).

When applied specifically to HSC operator modeling, which is the problem at hand, these techniques fall short. They all either require known patterns to analyze incoming data or are exclusively used as post hoc analysis tools. Additionally, all have little predictive power in a sequential, dynamic environment such as multiple operators controlling multiple highly autonomous UVs.

### ***2.2.2 Bayesian Methods***

The previous methods all belong to the family of statistical learning techniques in that they do not assume an a priori structure and instead rely solely on describing the human behavior as a set of stochastic

functions (Cucker & Smale, 2002). In contrast, Bayesian methods make the assumption of an underlying structure consisting of states and transitions between those states.<sup>1</sup> While this assumption restricts the particular form of the model, it also simplifies the formulation of the state space. Three commonly used Bayesian methods used for pattern recognition and prediction are partially observable Markov decision processes (POMDPs) (Sondik, 1971), hidden Markov models (HMMs) (Rabiner & Juang, 1986), and hidden semi-Markov models (HSMMs) (Guedon, 2003). Specifically applied to operator state modeling, POMDPs have proven to be successful for facial and gesture recognition (Sondik, 1971); but this domain is not generally cognitive in nature, and the goal of POMDPs is to compute the optimal policy given the current belief in the environment. HMMs and HSMMs, however, only focus on describing the most likely current belief in the environment and do not search for policy decisions. Additionally, HMMs have been shown to accurately classify and predict hand motions in driving tasks, which is a strong application of monitoring and prediction of sequential data (Pentland & Liu, 1995). In this work, however, the authors had access to the unambiguous ground truth linking the state of the model to the known hand positions. Thus, this method utilizes supervised training which makes it less useful for dynamic environments typical of HSC settings where the definitions of the states of the model are not known a priori (Boussemart, et al., 2009).

A recurring drawback of the previous techniques lies in their requirement of supervised training, a task that is labor intensive and may introduce bias via the human labeling process. However, recent research has shown success in HSC operator modeling with unsupervised learning of HMMs, as well as with model selection techniques that promote model generalizability (Boussemart & Cummings, 2008). Current work on HSMMs has been able to incorporate the temporal component of operator behavior, which was a limitation of using HMMs. It was also proposed and proven that HSMMs can both recognize normal operator behavioral patterns and flag abnormal patterns and events (Huang, 2009).

### ***2.2.3 HSMM-based Operator Modeling***

It has been hypothesized that the structure of HSMMs makes them well suited for modeling human behavior (Huang, 2009). Structurally, HSMMs consist of a set of hidden states, observable events, sojourn time, and transition probabilities. In Figure 2.4, each  $S_i$  is a hidden state, each  $a_{ij}$  is a transition probability between hidden states, the observable probability density functions  $P(o/S_i)$  link the state and the observable events each hidden state emits, and the sojourn probability density functions  $d_i(u)$  describe the amount of time an operator is likely to stay in each state.

---

<sup>1</sup> In this context, a state is defined as the set of partitions of the behavioral space that most likely explain the behaviors seen in the training data.



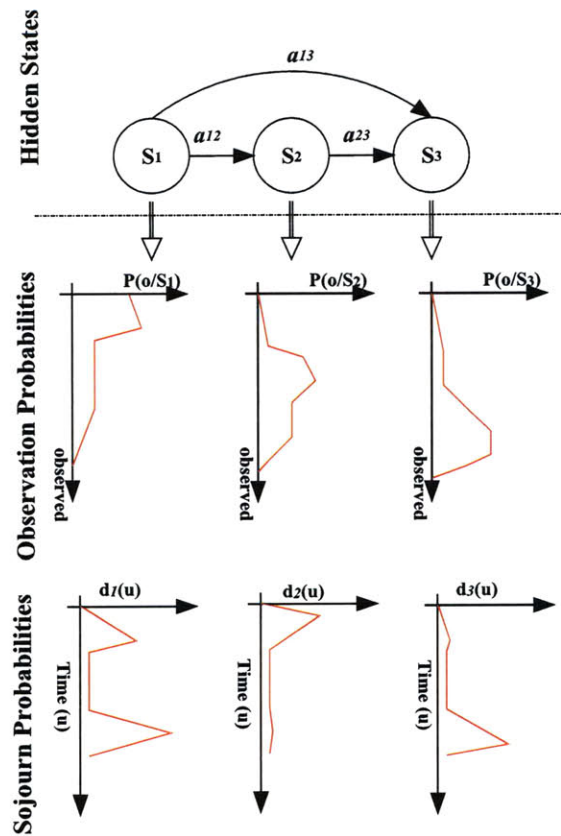


Figure 2.4: A three-state hidden semi-Markov model (Castonia, 2010)

Each hidden state emits observable events through a certain probability density function. Just as the hidden states in HSMMs must be inferred from observable events, cognitive states of an operator are not directly visible and must be inferred from his or her actions. Therefore, operator cognitive states can be likened to the hidden states in an HSMM, and operator interface interactions, i.e. mouse clicks, can be likened to the HSMM observable events. This structure allows for the prediction of future operator behavior but is limited by the semi-Markov assumption<sup>2</sup>, by the need for accurate training data in order to ensure the model accurately recognizes the correct patterns, and also by the decrease in confidence in future predictions as the prediction window increases.

<sup>2</sup> The semi-Markov assumption states that the transition to the next state is only based on the current state and the amount of time that has been spent in that state.

The complex probabilistic nature of the outputs from these models, and all Markov-based models, results in difficulty in designing a decision support interface for a team supervisor unlikely to possess the background required to understand them. This issue is discussed in the following section.

### ***2.3 Display of Probabilistic Data***

It is highly unlikely that a team supervisor in a complex HSC setting, such as the futuristic environment previously described, has the background in probability and modeling needed to understand the raw output from an HSMM-based DST. This is a concern because it has been well established that humans are often biased and subjective when given even simple probabilistic values (Tversky & Kahneman, 1974). Therefore, giving a HSC supervisor the raw probability density functions resulting from a HSMM-based model will likely result in confusion and may result in a negative effect on decision making, especially in time-critical scenarios. Also, the framing and context within which a probability is given to an individual may result in strong optimistic or pessimistic biases for the exact same probability (Tversky & Kahneman, 1981; Weinstein & Klein, 1995). The way in which probabilistic information is displayed, even as simple percentages in deciding whether or not to keep a set amount of money or take a percentage risk for an increased payout, has been shown to significantly affect what decision is made with that information (De Martino, et al., 2006). In a complex monitoring environment, subjects have been shown to completely ignore changes in a priori probabilities of subsystem failures (Kerstholt, et al., 1996). Unfortunately, these types of biases are difficult to remove (Weinstein & Klein, 1995), and thus extreme care must be taken to ensure that probabilistic data is displayed in a manner that reduces the tendency of operators to succumb to these biases.

The information must also be displayed in such a manner that that the supervisor neither blindly follows the algorithm's predictions and recommendations (misuse) nor fully disregards the algorithm's output (disuse) (Dzindolet, et al., 2003; Muir & Moray, 1996; Parasuraman & Riley, 1997). One technique to combat these biases is to provide the supervisor with information about how confident the algorithm is in regard to the given prediction (Bisantz, et al., 2000). This technique is important because the time sensitive nature of a HSC scenario will often require the supervisor to settle for a satisficing strategy as opposed to being able to evaluate all possible options, which will result in decisions being made with incomplete knowledge of the full situation (Gigerenzer & Goldstein, 1996). Misuse or disuse of the algorithm's output could lead to negative effects on decision making in this type of situation and/or further automation bias (Cummings, 2004).

Previous work confirms it is critical that the information be displayed to the supervisor with minimal probabilistic information and in a way that is clear, reduces decision biases, includes confidence in predictions, and promotes quick decision making. Otherwise, the inaccurate and/or biased interpretation of a probabilistic value may result in an incorrect decision with potentially fatal consequences.

## ***2.4 Background Summary***

The complex HSC task of supervising a team of multiple, independent operators each controlling multiple, heterogeneous UVs produces a difficult team monitoring scenario. Supervisors of such systems could become quickly overloaded and perform poorly without the aid of automation. One form of assistance is a decision support tool. Specifically, recent advances in operator modeling through HSMM-based behavioral pattern recognition and prediction could provide the algorithms to drive a real-time supervisor's predictive display. However, the difficulty lies in the effective display of the probabilistic output from the HSMM-based operator model since humans have difficulty interpreting and acting on probabilities. This thesis seeks to fill the literature gap that exists at the critical intersection of these three domains: team monitoring, human behavior pattern recognition and prediction, and the display of probabilistic data. The next chapter addresses the design of the HSMM-based operator state decision support tool and the steps taken to address the issues raised in this chapter.



### 3. Design of a HSMM-based Operator State Decision Support Tool

After reviewing related work, a cognitive task analysis (CTA) was performed in order to generate requirements, both functional and information requirements, for the task of supervising teams of 2-6 operators, each independently controlling a set of highly autonomous UVs (Castonia, 2009). This team size is consistent with literature defining small groups as consisting of 3-7 personnel. The types of interaction that occur with groups of more than 7 people begin to exhibit organizational behavior, as opposed to team behavior (Levine & Moreland, 1998), which is beyond the scope of this thesis. Furthermore, these assumptions on team size are ecologically valid because current UV operations are mostly conducted in teams of three people (DoD, 2007a).

After the completion of the CTA, a standard systems engineering design process (Figure 3.1) was used to develop a prototype display. The dashed box in Figure 3.1 highlights the portion of the systems engineering design process, which was the focus of this research and normally occurs in the early aspects of the acquisition phase. The final design is described in this chapter, while the full design process is described in more detail in Castonia (2009).

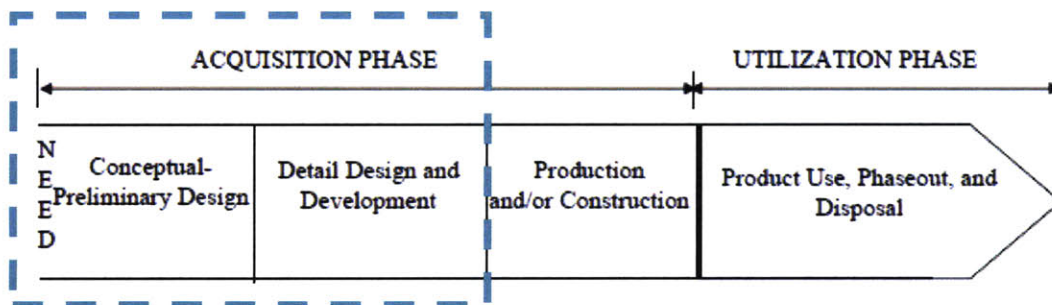


Figure 3.1: Systems engineering process (Blanchard & Fabrycky, 1998)

#### 3.1 Cognitive Task Analysis

A CTA is a tool used to analyze and represent the knowledge and cognitive activities needed in complex work domains (Schraagen, Chipman, & Shalin, 2000). Because it was determined that the main purpose of the DST would be to help supervisors determine if and when potentially detrimental problems could occur, the scope of the CTA is relatively narrow, focused primarily on the decision of whether to take action given a model prediction. Furthermore, the process of conducting a more traditional CTA requires access to subject matter experts, documentation, and predecessor systems to gain insight. However, these techniques are not applicable in a futuristic scenario, such as the scenario proposed in this work, where predecessor systems do not exist. The Hybrid CTA (Nehme, et al., 2006) addresses the shortcomings of a

traditional CTA when dealing with futuristic systems, but given the limited focus of this effort, only decision ladders were used to descriptively model critical decision processes and generate the functional and information requirements. The end result of this CTA is a display requirements checklist (Table 3.1) that details 12 requirements that should be met given DST prototypes with predictive capabilities. Conducting this, and all CTAs, is an iterative process that becomes more refined as more information about the situation becomes available.

### 3.1.1 Decision Ladders

Decision ladders break decisions down into the important information and information-processing activities required to make a decision. They also help a designer to see the three levels of human behavior hierarchies that exist within a set of tasks: skill-based, rule-based, and knowledge-based behavior (Rasmussen, 1983). Figure 3.2 shows a generic decision ladder with these hierarchy levels.

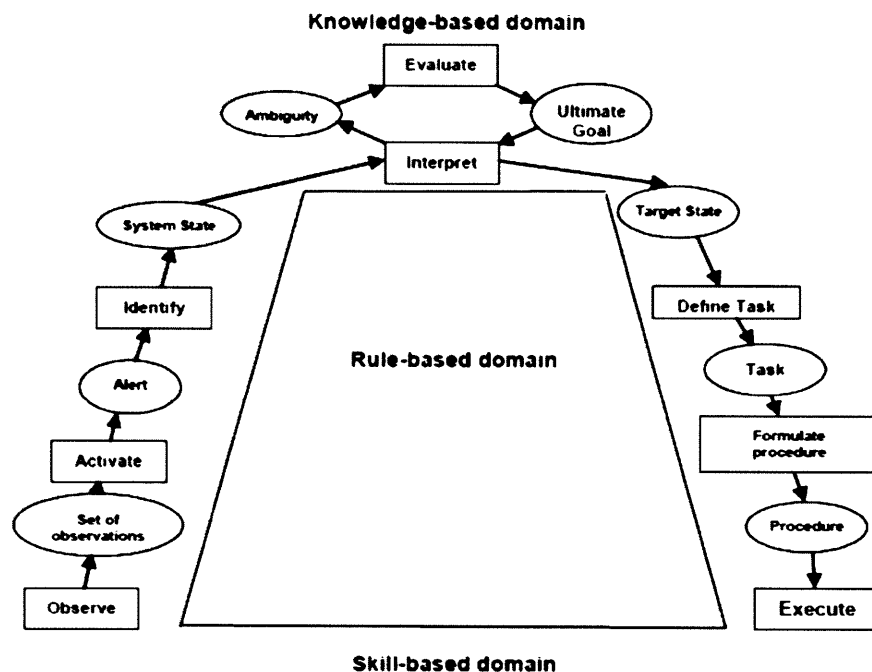


Figure 3.2: A decision ladder with human behavior hierarchies shown (Rasmussen, 1983)

The skill-based level relates to activities that take place without conscious control, which are typically behaviors that have reached a level of automaticity, such as tracking between lane markers while driving. The rule-based level relates to activities that rely on stored rules from prior experience in similar situations, such as looking for yellow or red text to appear on a screen when an alert sounds. The

knowledge-based level relates to higher level decision making resulting from the goals and cues from the environment, such as deciding what UV should be assigned to an emergent target. Each decision ladder includes boxes that portray the information processing activities and ovals that portray the information or knowledge that is produced from the information processing activities.

Given the focus of this research, which is to develop a predictive DST that alerts a supervisor to a possible problem for one or more team members, there are two fundamental decisions:

1. "Is there a problem?"
2. "What should I do to solve the problem?"

A decision ladder was created for each of the two critical decisions, and then functional and information requirements for the DST were derived from the decision ladders. In keeping with the hybrid CTA approach, each decision was represented by two decision ladders (an initial ladder and one augmented with display requirements), provided in Appendices A and B, respectively. These two decisions are detailed below.

*After receiving a visual and/or auditory alert to reference the DST for a possible problem (as determined by the HSMM-based operator model), the first critical decision the supervisor faces is determining if there is actually a problem. First, the source of the alert must be identified. In order for this to occur, the DST should display the information source that caused the alert so that it is readily apparent to the supervisor. The supervisor must then determine if operator actions caused the alert and if so, which actions. Consequently, the DST should display both what the model infers as the most likely cause of the alert and recent major environmental events, such as emergent targets, visual tasks, etc. that could have triggered the alert. After determining what caused the alert, the supervisor must then decide whether the alert requires action. To aid in this decision, the DST should display the prediction of future model performance to allow the supervisor the ability to develop appropriate confidence in the automation. If the supervisor decides the alert does not require action, then the DST must allow the alert to be easily ignored or reset. If the supervisor decides to change alert threshold levels, then the DST must ensure this is a quick process so that the supervisor can quickly return to monitoring the rest of the team. If the supervisor decides to intervene, then a procedure must be formulated by moving into the "What should I do to solve the problem?" decision ladder (Appendix A).*

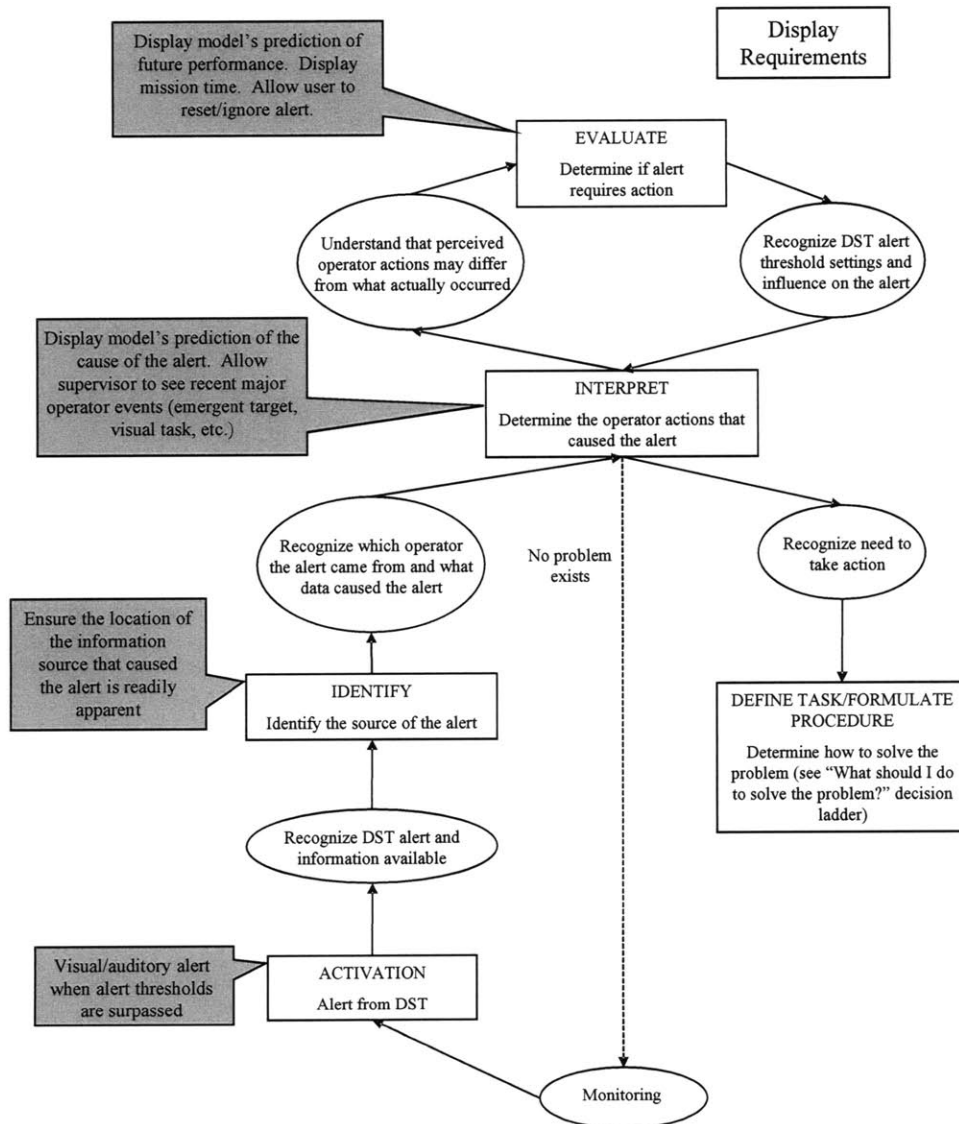
*The second critical decision involves the supervisor determining what must be done to solve the problem. Optimizing the time and frequency the supervisor intervenes are key factors in optimizing team performance (Brewer & Ridgway, 1998; Irving, et al., 1986; Mitchell, et al., 2004). The supervisor must identify possible solutions, such as advising the operator to change tactics, offloading a target and/or UV to another operator, obtaining an additional UV from another operator, changing alert threshold levels, or any combination thereof. This process should be aided by a DST that allows the supervisor to select the time frame of the data that is shown, including past operator performance levels. This would allow the supervisor to hone in on the information he/she desires. The supervisor is assumed to have the ability to view the operator displays, either remotely or in situ, in order to obtain more detailed information about current operator resources and actions as well as communicate with the operators.*

Throughout this entire decision-making process, the supervisor must be aware of how predictions change with passing time. As time passes, the problem may no longer require intervention because the operator may have adjusted his/her behavior accordingly, targets previously identified as hostile may now be confirmed as neutral, etc. Alternatively, the problem may quickly worsen and may require immediate intervention. In that case, the supervisor would have to make a decision quickly, and a less than optimal solution will suffice.

As mentioned earlier, each decision ladder was then used to derive display requirements. Each information processing activity, denoted by a box in the decision ladder, was assessed to determine which functional and information requirements would be necessary to allow the supervisor to obtain the information or knowledge that is contained in the corresponding oval. Figure 3.3 shows the display requirements derived from the first decision ladder. In this example, in order to recognize which operator an alert came from and what data caused the alert, the requirement is that the location of the information source that caused the alert be readily apparent. The completed display requirements for all decision ladders can be found in Appendix B. Note that the display requirements simply identify what information needs to be displayed, not how it will be displayed. The decision of how to display the information is addressed in the conceptual design phase.



**“Is there a problem?”**  
with display requirements for decision support tool



**Figure 3.3: Display requirements for the “Is There a Problem?” decision ladder**

### 3.1.2 Display Requirements Checklist

The tangible output from the CTA is the display requirements checklist provided in Table 3.1. It contains 12 different requirements that were obtained through the decision ladder display requirements. These requirements are divided into the problem identification and problem solving categories, which represent

the two primary functions of the DST. This checklist is designed to aid in the development and evaluation of prototype displays for team supervisor DSTs that provide predictions of future operator performance.

**Table 3.1: Display requirements for DSTs that provide predictions of future operator performance**

Type	Requirement Description
<b>Problem Identification</b>	<ol style="list-style-type: none"> <li>1. Alerts supervisor when alert thresholds are surpassed (visual and/or auditory)</li> <li>2. Location of alert information source is readily apparent at first glance</li> <li>3. Displays model's prediction of the cause of the alert</li> <li>4. Shows recent major operator events (emergent target, visual task, etc.)</li> <li>5. Displays model's prediction of future performance</li> <li>6. Allows user to reset/ignore alert</li> <li>7. Communication capability with operators (assumed to be met with other resources)</li> </ol>
<b>Problem Solving</b>	<ol style="list-style-type: none"> <li>8. Displays accuracy of model predictions of past events</li> <li>9. Displays history of operator interactions with UV interface</li> <li>10. Contains ability for supervisor to alter the view in order to obtain time-range specific data</li> <li>11. Displays operator identifying information</li> <li>12. Alert threshold levels are easily adjustable</li> </ol>

### ***3.1.3 Cognitive Task Analysis Summary***

The complex work domain of a supervisor and team of operators independently controlling multiple, highly autonomous UVs was analyzed through the CTA process described in this section. Specifically, decision ladders were used to generate the functional and informational requirements shown in Table 3.1. While the chosen operator model did not influence the output of the CTA, it greatly affected the way that the requirements generated by the CTA were implemented during the design phase, as described in the following section.

## ***3.2 HSMM-based Operator State Decision Support Tool***

Since the DST is designed as a secondary display in the supervisor's suite of displays, it is designed for a small, possibly portable display with industry standard SXGA resolution (1280 x 1024 pixels). The display requirements checklist in Table 3.1 was used to develop and revise each display revision. Figure 3.4 identifies the four major sections of the DST that will be discussed in this section.

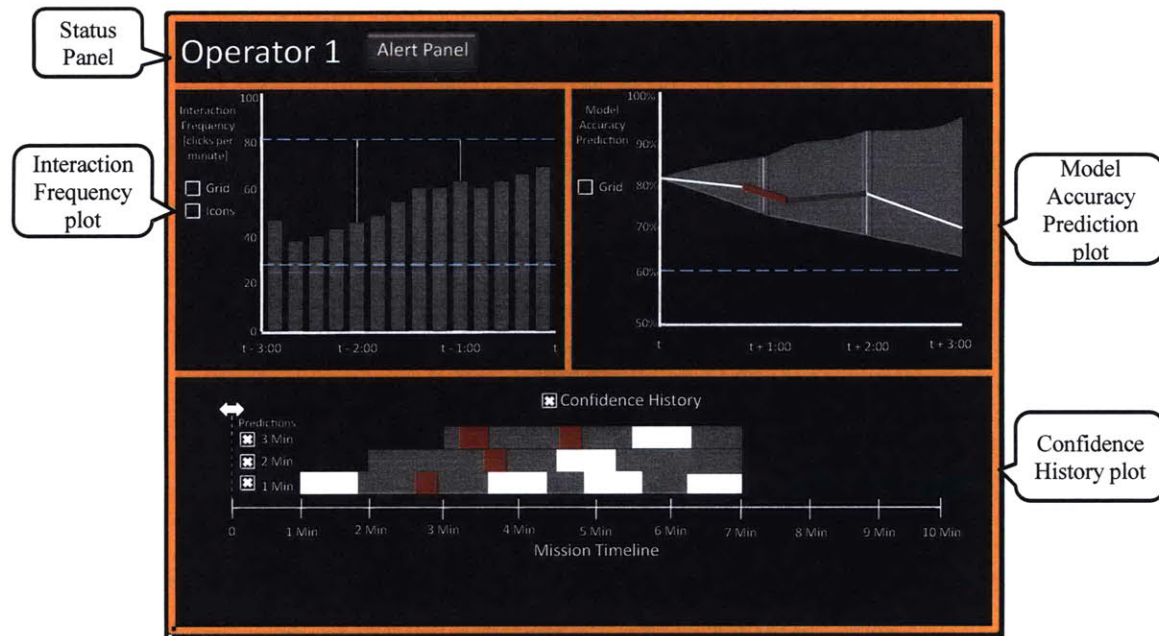


Figure 3.4: DST (Castonia, 2009)

### 3.2.1 Interaction Frequency Plot

The upper left hand corner of the DST contains the Interaction Frequency plot shown in Figure 3.5. Historical data for operator clicks per minute are shown as a way to give the supervisor an unprocessed way to infer operator workload (Maule & Hockey, 1993). The horizontal axis ranges from three minutes in the past ( $t-3:00$ ) to the current time ( $t$ ). The vertical axis ranges from zero clicks per minute to 100 clicks per minute.

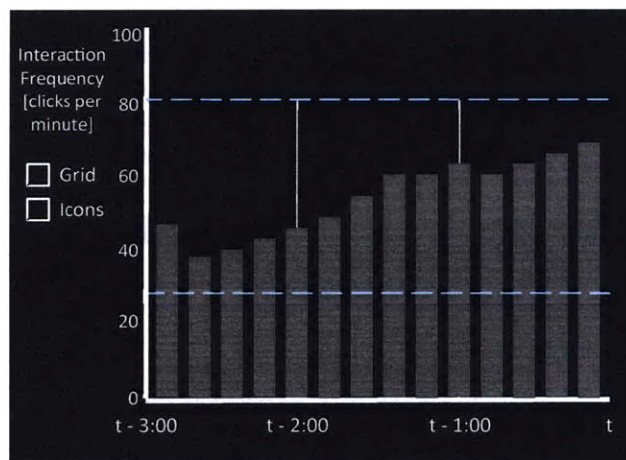
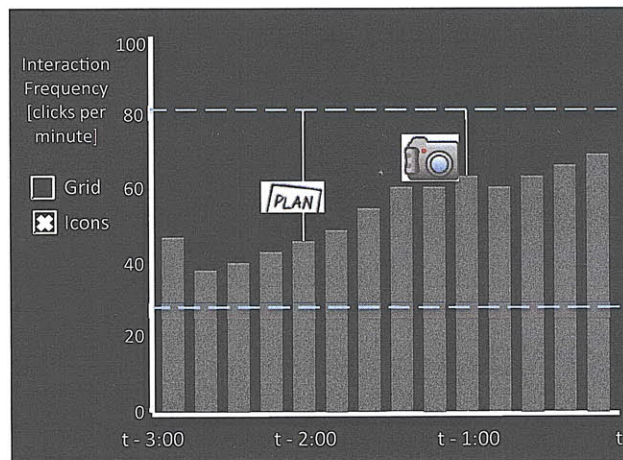
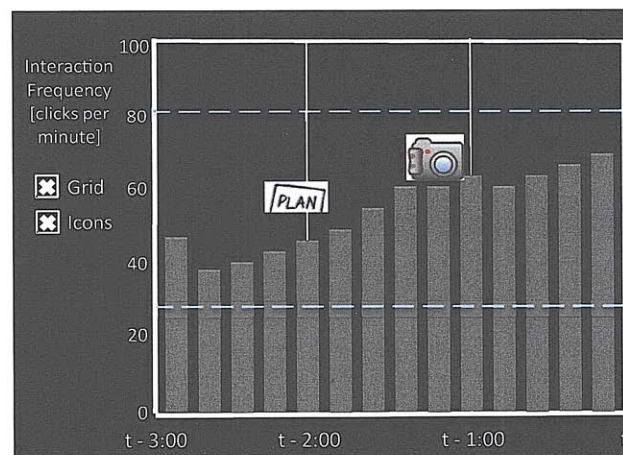


Figure 3.5: Interaction Frequency plot

One of the display options for the Interaction Frequency plot is the Icons selection box on the left side of the display. This toggles the display of icons on the Interaction Frequency plot that represent major operator events as seen in Figure 3.6. The use of icons immediately informs the supervisor to the possible cause of a recent increase/decrease in operator interaction frequency. Icons are expected to be useful in helping the supervisor to quickly determine if there is a problem and what must be done in order to solve the problem. Such time savings could be critical in time-pressed situations typical of command and control settings.



**Figure 3.6: Interaction Frequency plot with icons shown**



**Figure 3.7: Interaction frequency plot with icons and grid shown**

In an attempt to increase the data-ink ratio (Tufte, 1983), the proportion of the display's ink devoted to the non-redundant display of data-information, unnecessary portions of grid lines and plot outlines are

removed. However, the Grid selection box on the left side of the Interaction Frequency plot allows the supervisor to turn these grid lines and plot outlines back on if he/she wishes (Figure 3.7).

### 3.2.2 Model Accuracy Prediction Plot

The upper right hand corner of the DST in Figure 3.4 is the Model Accuracy Prediction plot and is the portion of the display that most utilizes the HSMM-based operator model. The HSMM-based model generates transition probabilities between different system states for each instance in time and utilizes those likelihoods, combined with the observed operator actions, to predict the future accuracy of the model and to predict the likelihood of future operator behavior being “normal.” The HSMM-based model also provides a prediction as to which operator behavior was deemed to be anomalous by identifying the discrepancies between the expected state transitions and the observed operator actions. This model provides the team supervisor with information about how “normal” the operator behavior is predicted to be, as well as the predicted cause of any alerts that arise from “abnormal” behavior. As a reminder, the HSMM-based model cannot determine whether the abnormal behavior is detrimental.

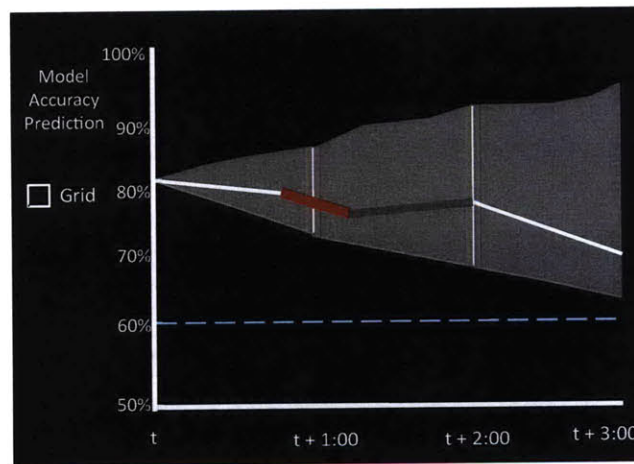


Figure 3.8: Model Accuracy Prediction plot

The Model Accuracy Prediction plot (Figure 3.8) incorporates predictions about each future predicted state and the expected duration inclusive of the next three minutes. The horizontal axis ranges from the current time (t) to three minutes in the future (t+3:00). The vertical axis ranges from 50% model accuracy to 100% model accuracy. If the model accuracy prediction is low, then the model is expected to not be able to accurately predict operator behavior and the operator behavior is considered “abnormal.” If the model accuracy prediction is high, then the model is expected to predict future operator behavior accurately and the operator behavior is considered “normal.” While coding the DST, Huang (2009)

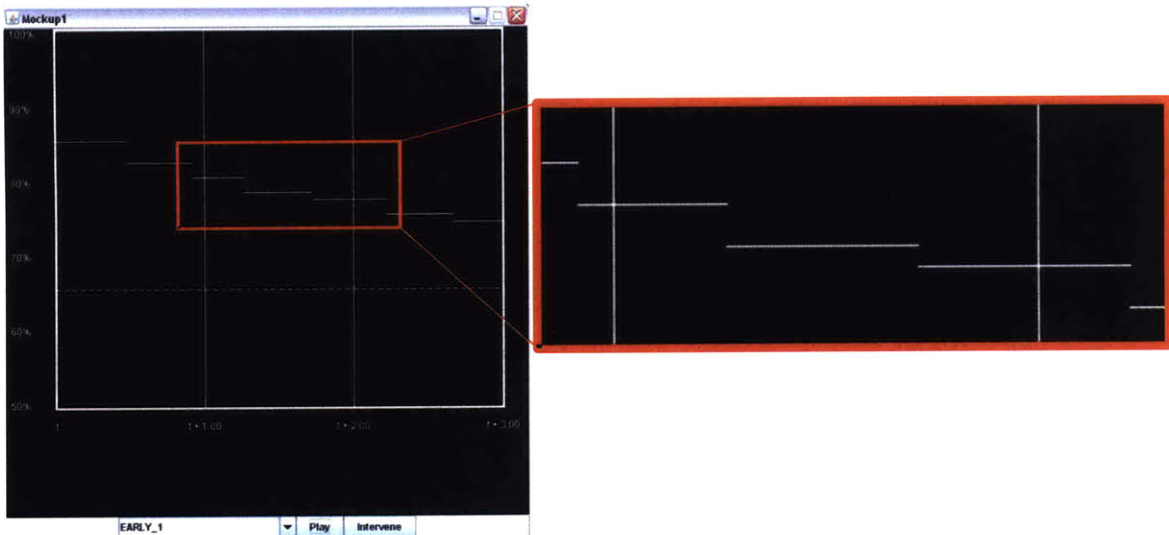
showed that the output of the HSMM-based operator model would require a scoring metric, the Model Accuracy Score (MAS), that represents the current model accuracy based upon the quality and timing of previous predictions. Hypothetical future operator actions are then appended to the current model state and the appropriate future MAS predictions are populated. The MAS takes on values from 50-100, which presents the supervisor with feedback on future prediction accuracy.

The range of possible future model accuracy values, corresponding to the different possible state transitions patterns, is represented by the gray shaded area in Figure 3.8. As expected, the range of possible values, and thus the area of the gray shaded region, increases as the model predicts farther into the future. The white/red/dark gray line represents the midline of the range distribution. When this line drops below a specified threshold, shown by the blue dashed line in Figure 3.8, an alert is triggered. This allows the supervisor to easily see how close the Model Accuracy Prediction is to causing an alert.

While the range and midline of the distribution of model accuracy values are important, so is the confidence the model has in those prediction. This information is available through the color of the midline (Figure 3.8). Each color is mapped to a level of prediction quality: high (white), medium (gray), and low (red). The specific quantities used to classify these three levels depend on the underlying models and supervisor preference. For the given prototype, expected deviations of less than five percent from the midline were mapped to high confidence, expected deviations of five to ten percent from the midline were mapped to medium confidence, and expected deviations of greater than ten percent from the midline were mapped to low confidence. The colors were chosen to reinforce the mental model of white = definite/confident, gray = uncertainty, and red = warning, that is consistent throughout the display, and to ensure the DST can be used by supervisors with red-green color blindness. This color scheme is important since approximately 10% of the male population is color-blind, and 95% of all variations in human color vision involve the red and green receptor in human eyes (Gegenfurtner & Sharpe, 2001). Since an alert is generated when the midline drops below a supervisor-dictated threshold and it is expected that the supervisor will immediately want to know the prediction quality of this alert, this mapping of prediction quality to midline color will result in minimal search time to obtain the relevant data.

While the design for the midline of the model accuracy distribution has always been linear, the actual output from the HSMM-based model is discontinuous horizontal lines. The model output gives the expected amount of time the operator will spend in the current state, the probabilities of moving to each of the other states, and the expected amount of time the operator will spend in each of those states. Therefore, each horizontal line in Figure 3.9 represents the most probable next state, where the length of

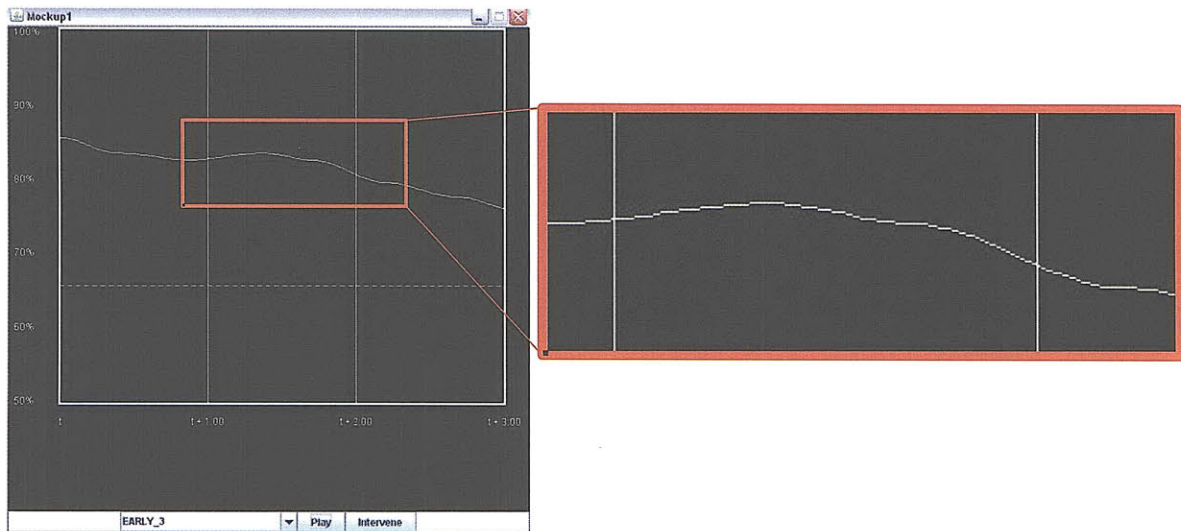
each line is the predicted length of time the operator will stay in that state. However, a preliminary experiment was conducted that showed users reported higher understanding and were more confident in the HSMM-based data presented to them in a curvilinear representation (Figure 3.10 or Figure 3.11) than in the discrete representation (Figure 3.9) that is most suitable to the actual model output (Castonia, 2009). Therefore, the DST in Figure 3.4 incorporates a curvilinear representation of the HSMM-based data.



**Figure 3.9: Discrete variation of the Model Accuracy Prediction display (Castonia, 2009)**



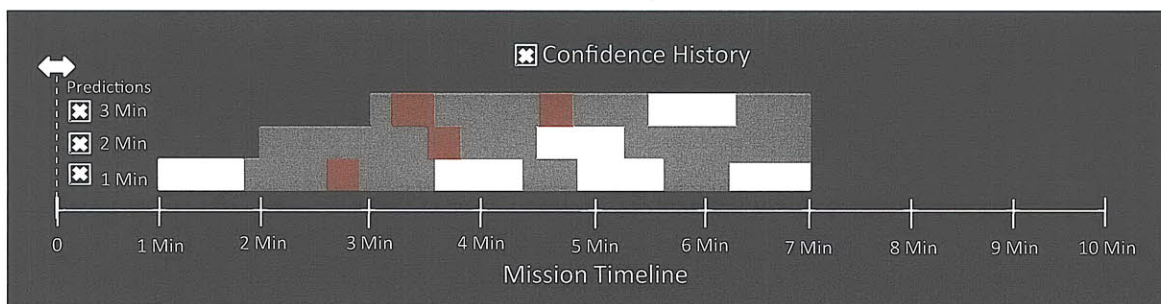
**Figure 3.10: Linear variation of the Model Accuracy Prediction display (Castonia, 2009)**



**Figure 3.11: Curved variations of the Model Accuracy Prediction display (Castonia, 2009)**

### 3.2.3 Confidence History Plot

If alerted to a drop below the set threshold in the Model Accuracy Prediction plot, the supervisor would likely want to know how well the model has been performing throughout the mission. If the model has been performing well, then the supervisor may lead toward intervening in the scenario. However, if the model has been performing poorly, the supervisor may be more inclined to reset the alert. In order to provide the supervisor with information about model performance over time, a high level view was created as shown in Figure 3.12.

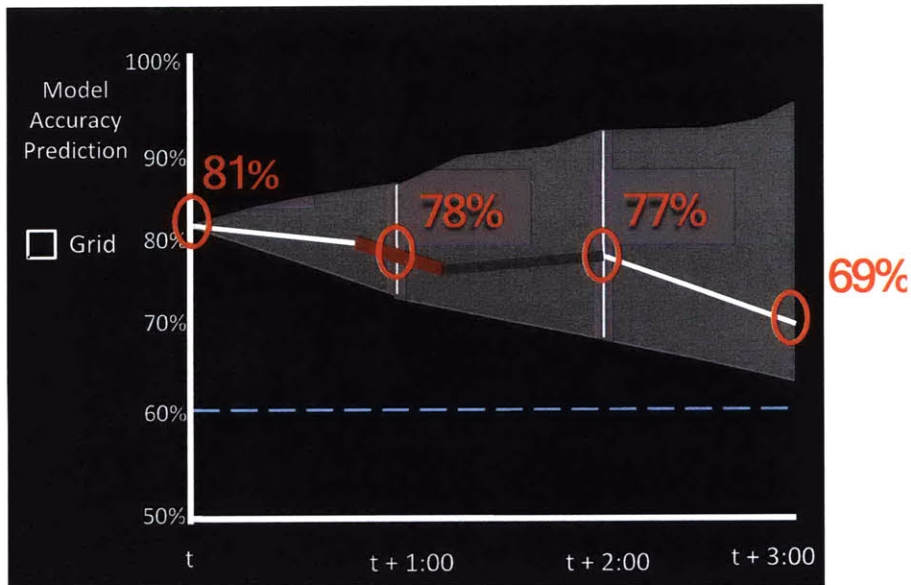


**Figure 3.12: Confidence History plot**

To create this display, the historical Model Accuracy Predictions are compared to the observed model accuracy for each point in time. This comparison allows the supervisor to quickly and confidently identify how accurate the model has been at making predictions over the course of the entire mission, influencing the decision of whether to trust the model's predictions and actually intervene when there is an alert, or to



simply ignore it. For each instant in time, the value of the Model Accuracy Prediction midline for the one, two, and three minute predictions are stored and then compared to the observed model accuracy at the time when those predictions come to fruition.

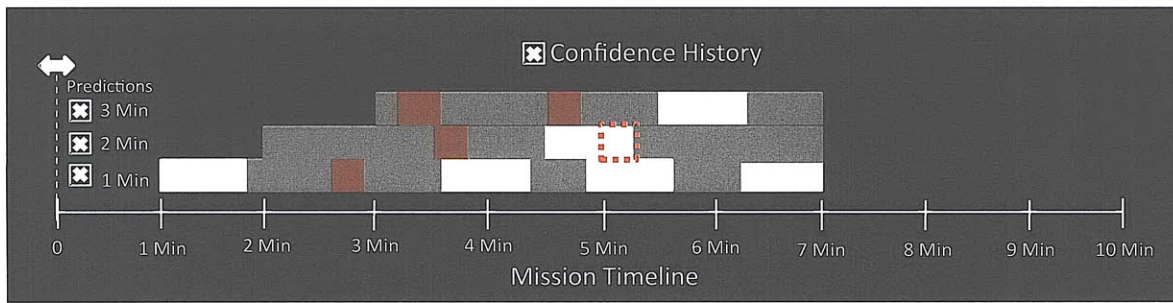


**Figure 3.13: Model Accuracy Prediction plot with applicable Confidence History plot values annotated**

An example of the applicable values from the Model Accuracy Prediction plot that are used to produce the Confidence History plot can be seen in Figure 3.13. In this example, the current observed model accuracy is 81%, the one minute future prediction is 78%, the two minute future prediction is 77%, and the three minute future prediction is 69%. These values are circled in Figure 3.13. If the current mission time in Figure 3.13 is 3 minutes ( $t = 3$  minutes), then we see that the Model Accuracy Prediction for a mission time of 5 minutes ( $t + 2:00$ ) is 77%. Two minutes later, at a mission time of 5 minutes, assume that the observed model accuracy is 80%. These two values would then be compared to populate the Confidence History plot because 77% was the two minute prediction for a mission time of 5 minutes and 80% was the observed model accuracy for a mission time of 5 minutes. The difference is 3%, which describes how good the model did at predicting two minutes into the future. A smaller difference is indicative of more accurate predictions, while a larger difference is indicative of less accurate predictions.

The difference between the Model Accuracy Prediction for a specific mission time and the observed model accuracy for that specific mission time is the information contained in the Confidence History plot. This information is color coded in a manner similar to the three levels of Prediction Quality: deviations of

less than five percent correspond to high confidence (white), deviations of five to ten percent correspond to medium confidence (gray), and deviations of greater than ten percent from correspond to low confidence (red). Therefore, the example just described would result in a white bar (because the deviation between the predicted and observed values was 3%) starting at the 5 minute location on the horizontal axis (the mission timeline) and the 2 minute location on the vertical axis (the prediction axis), as annotated by the highlighted portion of Figure 3.14. That bar would stay white and continue to populate the plot to the right until the deviation between the two minute Model Accuracy Prediction midline and observed model accuracy reaches 5% or greater. This appears in Figure 3.14 at a mission time of about 5:20 where the deviation between the two minute Model Accuracy Prediction and observed model accuracy increased to the medium confidence range, thus changing the bar to gray.



**Figure 3.14: Annotated Confidence History plot**

These colored bars allow the supervisor to quickly determine how confident he/she should be in the current predictions, which has been shown to positively influence decision maker accuracy (Bisantz, et al., 2000). If the Confidence History plot is mainly white, the model’s predictions of operator behavior have been accurate and the supervisor should have high confidence in the current predictions contained in the Model Accuracy Prediction plot. If the plot is mainly red, the model’s predictions of operator behavior have not been accurate and thus the supervisor’s confidence in current predictions contained in the Model Accuracy Prediction plot should be low.

The supervisor has several options when using the Confidence History plot. First, the entire plot can be toggled on/off. This toggle allows the supervisor to focus on the Model Accuracy Prediction and Interaction Frequency plots when he/she is not concerned about the historical model performance. Additionally, the supervisor is able to manipulate what information is shown in the plot through the use of the selection boxes on the left side of the display and the click-and-drag movable vertical axis. The selection boxes correspond to the one minute, two minute, and three minute predictions from the Model

Accuracy Prediction plot and allow the supervisor to view any combination of the three that he/she wishes. This is useful if, for example, the three minute predictions have been very inaccurate and the supervisor has decided not to take them into account any longer. Clicking the corresponding toggle box removes the confidence history for the three minute predictions, as seen in Figure 3.15, and allows the supervisor to focus solely on the confidence history for the one and two minute predictions. The click-and-drag movable axis, highlighted in Figure 3.15, allows the supervisor to hide old data. This is useful if the mission lasts an extended period of time and the supervisor is no longer concerned about how accurate the model was much earlier in the mission. Overall, the Confidence History plot provides the supervisor with a flexible way to evaluate past model performance in order to aid in the important decision of when to intervene.



**Figure 3.15: Confidence History plot with 3 min prediction not shown**

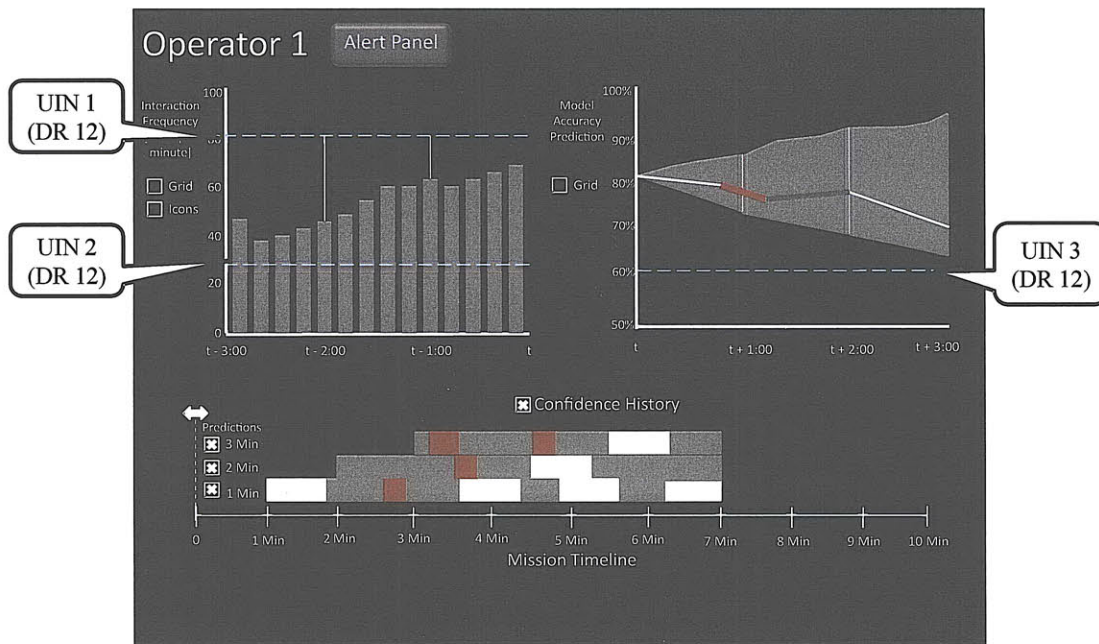
### 3.2.4 Status Panel

The status panel, shown in Figure 3.16, is at the top of the DST and provides the supervisor with the operator name, access to the alert panel, displays alerts, and provides the interface for the supervisor interaction once an alert has occurred. The operator name is provided so that the supervisor may utilize customized alerts that may have been saved to that particular operator from a previous mission and also so that the supervisor may recall prior experiences that he/she has with that operator in order to influence the decision making process throughout the mission.



**Figure 3.16: Status Panel**

As mentioned during the CTA in Section 3.1, it is assumed that the supervisor will be focused on other displays and tasks until alerted to reference the DST. These alerts are generated when preset alert thresholds, shown in Figure 3.17, are surpassed.



**Figure 3.17: DST with UIN's highlighted**

The requirement that these thresholds are easily adjustable applies not only to specific alerts, but also for the situation where the supervisor may want different thresholds for different operators, i.e. novices might have more narrow ranges and experts might have larger ranges of expected behaviors. In order to provide this functionality, all alert thresholds are designed as user-initiated notifications (Guerlain & Bullemer, 1996) that can be adjusted directly on the interface by a click-and-drag function or through accessing the alert panel. This allows the supervisor to quickly and easily set different user-initiated notification (UIN) threshold levels for each operator.

Each UIN threshold line is represented as a blue dashed horizontal line and the description of each UIN in Figure 3.17 is given in Table 3.2. When the interaction frequency drops below the lower line (UIN 2) or rises above the upper line (UIN 1), then the DST will display an alert similar to the alert in Figure 3.20. The UIN threshold line on the Model Accuracy Prediction plot (UIN 3) represents the lowest predicted model accuracy the supervisor is willing to tolerate before being alerted. When the midline of the model

accuracy prediction drops below this line, the supervisor is alerted and the predicted cause of the alert is displayed.

**Table 3.2: UIN Descriptions**

	<b>Plot Location</b>	<b>UIN Description</b>
UIN 1	Interaction Frequency	Alerts supervisor when operator is clicking more often than the set value.
UIN 2	Interaction Frequency	Alerts supervisor when operator is clicking less often than the set value.
UIN 3	Model Accuracy Prediction	Alerts supervisor when the midline of the model accuracy prediction drops below the set value.

The Alert Panel, shown in Figure 3.18, is accessed through the Alert Panel button on the Status Panel. The supervisor may use the Alert Panel to set the parameter, threshold value, alert window, recurrence (once or always), and alert type (visual, auditory, or both) for as many UINs as he/she wishes. The settings for the three UINs shown throughout this chapter and specifically in Figure 3.17 can be seen in the bottom of the Alert Panel in Figure 3.18. Selecting one of these existing alerts and clicking Edit allows the supervisor to change any of the settings, such as changing the alert window for the Model Accuracy Prediction UIN from ‘Present - +3 minutes’ to ‘Present - +2 minutes’ because of existing low confidence in the three minute predictions. The supervisor can also change the recurrence from ‘Once’ to ‘Always’ so not to have to create a new UIN once the corresponding alert is triggered. The UIN may also be reset by clicking the Acknowledge/Reset Alert button on the main display as seen in Figure 3.19.

The Alert Panel also allows the supervisor to set alert thresholds for system events such as emergent targets, camera/visual tasks, UV losses, neutralization of targets, etc. It is expected that users would initially rely primarily on preset alerts, while allowing expert users to tailor the DST to his/her preference, even in real-time.

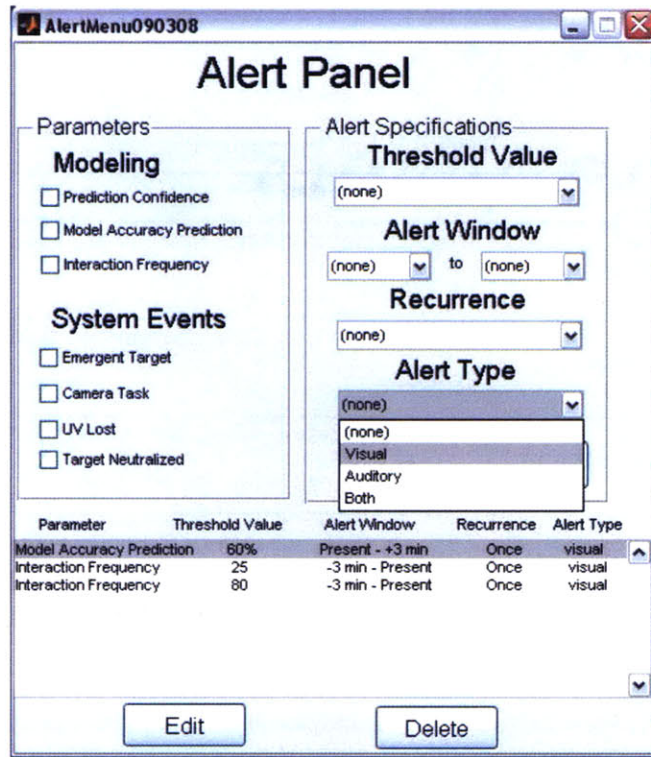


Figure 3.18: Alert panel

Each alert has a corresponding alert message. The alert message is displayed on the Status Panel as seen in Figure 3.19. The message contains the model’s prediction of what caused the alert and is kept succinct in order to allow the supervisor to quickly scan the alert and understand its meaning. The alert message in Figure 3.19 states that the alert has likely been caused by Operator 1 neglecting the underwater unmanned vehicles (UUVs) and favoring the high altitude long endurance unmanned vehicles (HALEs).



Figure 3.19: Status Panel with alert shown

After each alert, the supervisor must decide whether to intervene in the situation or reset the alert. The Intervene button is designed to interact with other systems to allow the supervisor to implement the chosen solution. The Acknowledge/Reset Alert button is used if the supervisor determines that the alert does not require intervention. When the Acknowledge/Reset Alert button is clicked, the alert message is removed from the Status Panel and the alert is reset. If the alert was set for an occurrence of ‘Once,’ this results in the removal of the corresponding blue dashed UIN line. If the alert was set for an occurrence of

‘Always,’ the corresponding blue dashed UIN line will remain on the DST and the supervisor will be alerted if the line is crossed in the future.

### 3.3 Summary

The DST brings together the information required to make the two fundamental decisions identified in the CTA for the task of team supervision of 2-6 operators, each independently controlling a set of highly autonomous UVs.

1. “Is there a problem?”
2. “What should I do to solve the problem?”

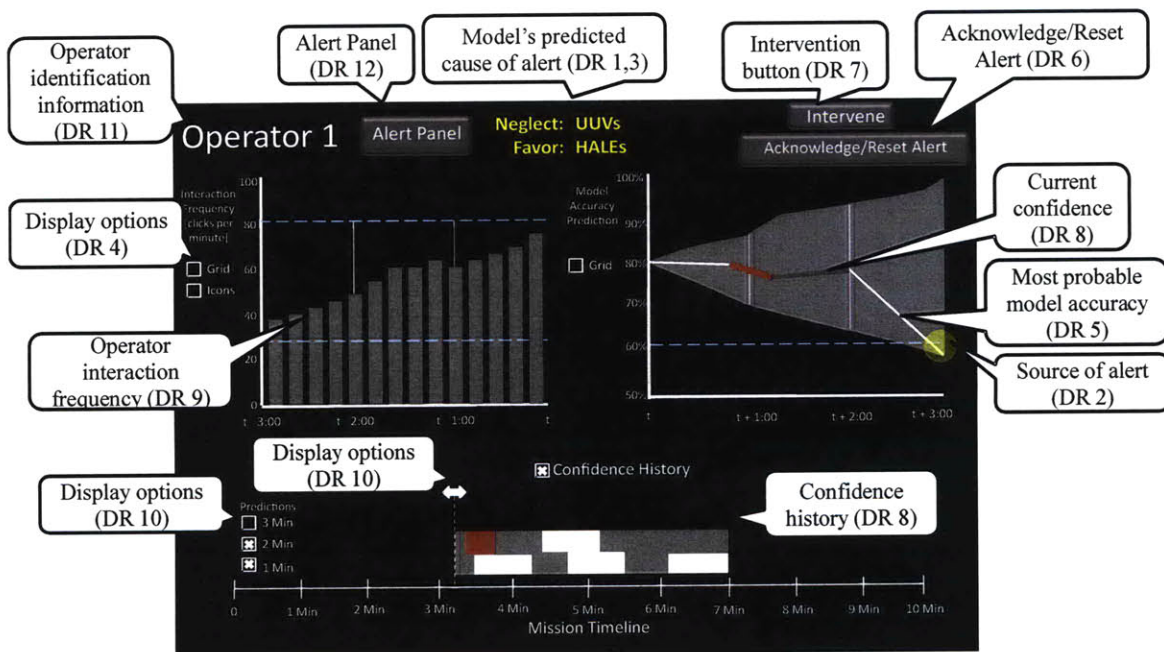


Figure 3.20: Annotated DST with Design Requirements (DR) shown (Castonia, 2009)

Furthermore, each Design Requirement (DR) listed in Table 3.1 is met by at least one feature on the DST and there are not any features on the DST that do not meet at least one Design Requirement, as seen in Figure 3.20. The callouts on Figure 3.20 identify different features of the design as well as the DR from Table 3.1 met by each feature. The design was developed with the goal of displaying the necessary information in the simplest method possible, in order to ensure the supervisor can quickly and easily find whatever he/she needs to answer the questions posed above.

Although the DST is designed with UV operations in mind, the design is applicable to any HSC scenario given a working model of current and predicted operator behavior. Additionally, since the DST is a unique design for a futuristic application, formidable testing must be completed in order to evaluate its effectiveness. The next chapter describes a human subject evaluation that was completed in order to accomplish this task.



## **4. Experimental Evaluation**

After the completion of the development process described in Chapter 3, an experiment was designed to test the effectiveness of the resulting DST. Specifically, the experiment was designed to address the problem statement in Chapter 1, which is determining whether team supervisors with an HSMM-based DST perform better than supervisors without the DST in terms of correctly identifying current and future problems, solving current problems efficiently and accurately, correctly preventing future problems from occurring, and limiting unnecessary interventions.

The experiment involves a supervisor whose job it is to monitor three operators. Each operator has an independent area of responsibility over which he/she must direct UVs to monitor pre-set and emergent targets, while avoiding dynamic threat areas. The team supervisor of the three operators must oversee the mission which, for the purposes of this experiment, is assumed to occur from a remote location. This simulates a supervisor in a command and control center possibly thousands of miles away from the operators, such as Air Force unmanned aerial vehicle pilots receiving commands from a supervisor who is in a remote command center. In this setting, the supervisor is able to view all three operator interfaces remotely and must make decisions of when to intervene in the different scenarios presented to him/her in order to best improve team performance.

In this experiment, the operators and all of the interactions with the UVs and targets within the operator interface were simulated. This was done to ensure consistency of test scenarios and control the variability in the data. Thus, the only actual humans in this experiment were team supervisors. Details about the participants, apparatus, procedure, and experimental design will be presented in this chapter.

### ***4.1 Participants***

Participants were recruited via email and word of mouth. Of 30 total participants, 21 were male and 9 were female. The average age was 19.73 years with standard deviation of 1.20 years and a range of 18-26 years. Of the 30 participants, 29 were MIT students (27 undergraduate, 2 graduate).

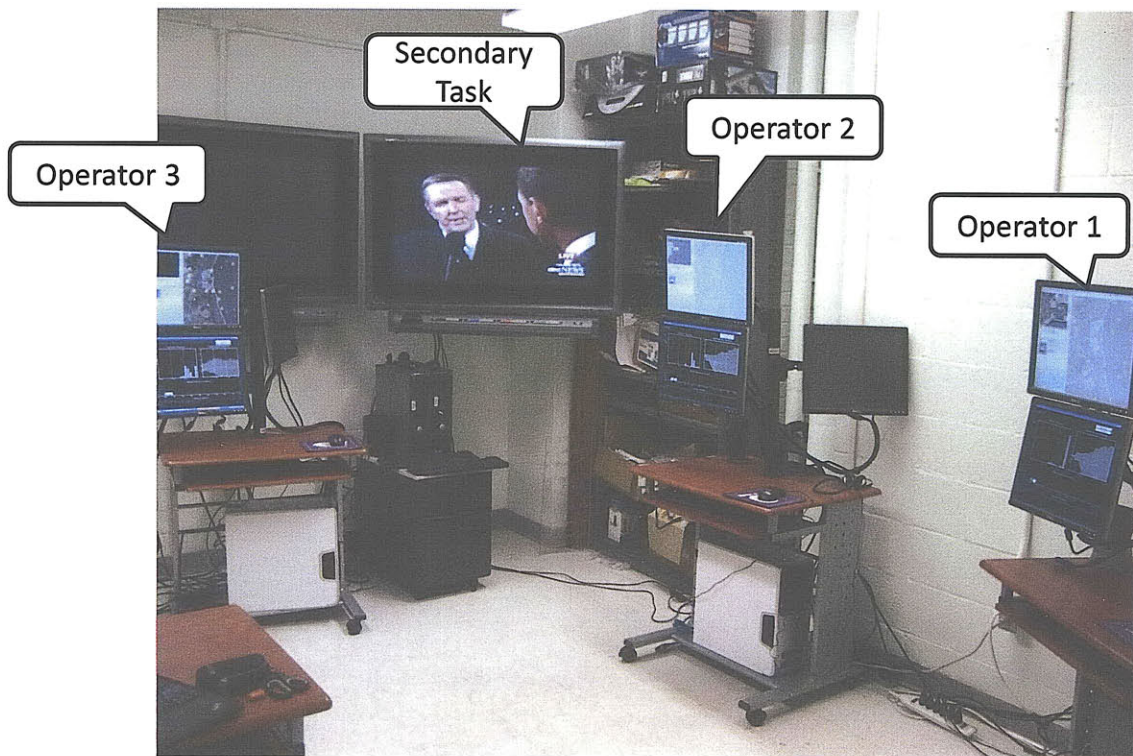
### ***4.2 Apparatus***

The experimental environment consisted of three different displays: the operator interface, the team supervisor interface, and the secondary task display. Each operator interface and the team supervisor interface were displayed on 17-inch Dell TFT LCD monitors connected to a Dell Dimension tower

containing a Pentium D 2.80GHz CPU and 2.00 GB ram. The secondary task was shown on a 42-inch (1024x768 pixels) wall-mounted plasma screen. The overall testing environment layout will now be described, followed by the operator interface, team supervisor interface, and secondary task display.

#### *4.2.1 Testing Environment Layout*

The experimental team consisted of three operators (simulated) and a team supervisor. The testing environment can be seen in Figure 4.1. The three operator workstations were spread around the room so that it was easy for the experimenter to identify which operator the subject was paying attention to at any given moment. The subjects were allowed to move around as they wished in order to best supervise their team. The three operator workstations were started simultaneously over a network connection from a fourth computer to ensure all scenarios were shown to the subjects in the exact same manner.



**Figure 4.1: Testing environment layout**

Each operator station was configured as seen in Figure 4.2, with the RESCHU interface, discussed in the next section, displayed on the top monitor and the DST displayed on the bottom monitor. For the 15 subjects that were randomly selected to conduct the experiment without the DST, the bottom monitor was turned off.

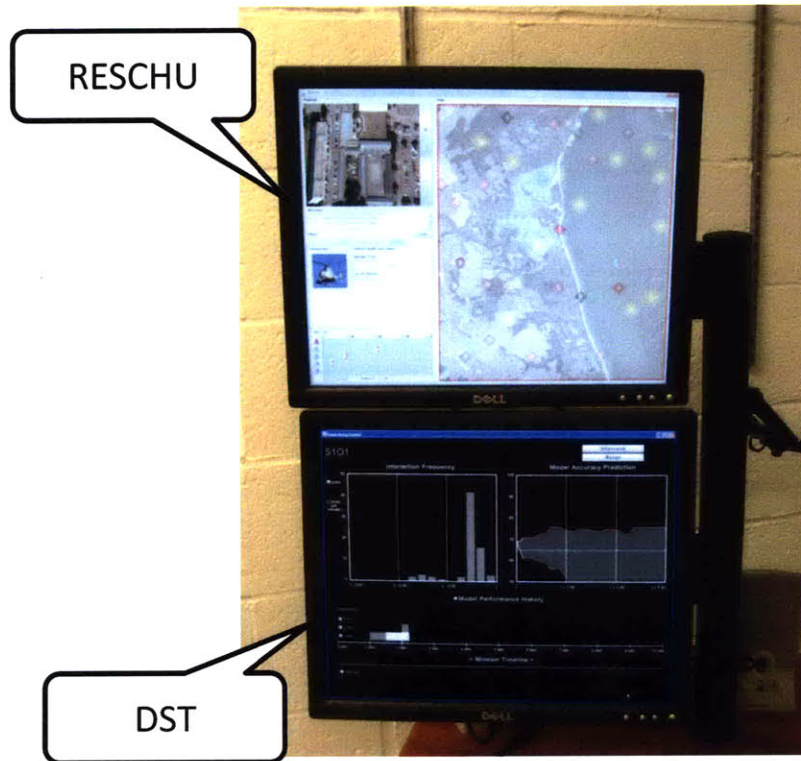


Figure 4.2: Operator workstation (RESCHU and DST)

#### 4.2.2 Operator Interface: *RESCHU*

The Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles (RESCHU) was chosen as the operator interface because of its functionality of allowing a single operator to control multiple, heterogeneous UVs in simulated intelligence, surveillance, and reconnaissance (ISR) tasks (Nehme, 2009). In RESCHU, it is the goal of each operator to visit as many targets as possible in order to correctly complete visual search tasks and obtain the maximum score. This requires the operator to dynamically plan and re-plan UV paths due to dynamic threat areas and emergent targets.

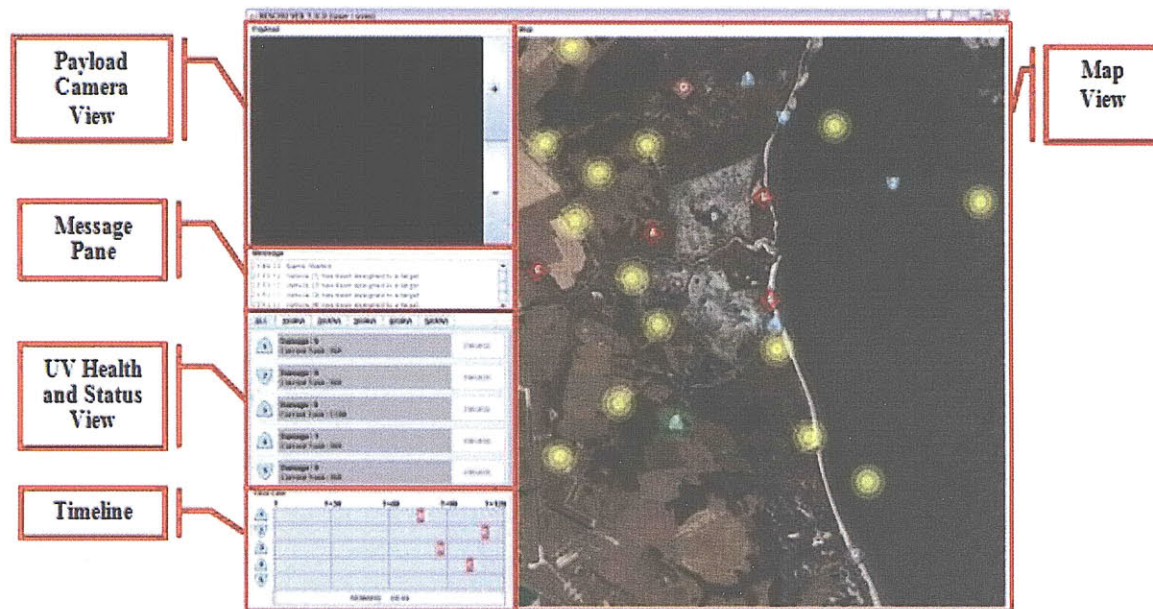


Figure 4.3: Annotated screenshot of the operator display – RESCHU (Nehme, 2009)

As annotated in Figure 4.3, RESCHU consists of five different components: a large map view, payload camera view, message pane, UV health and status view, and a timeline.

- The map view shows the locations of threat areas (yellow circles), UVs (blue), unknown targets (gray diamonds), and targets that need to be searched (red diamonds), as well as the current paths of the UVs (blue lines). The UVs with the green halo around them are the high altitude long endurance unmanned vehicles (HALEs) that must visit unknown targets before the other UVs may conduct the search task.
- The payload camera view, shown in Figure 4.4, is only engaged when a UV arrives at a target that needs to be searched (red diamonds). This view simulates the operator searching for an object, such as a white car, in the target area. If the operator successfully identifies the correct object, his/her score is increased by one. At the completion of the visual search task, the target disappears and the operator may once again guide the UVs to remaining targets.
- The message pane logs system messages, such as when UVs arrive at targets or what the operator needs to search for in the payload camera view, for quick reference by the operator.
- The UV health and status view provides information such as how much damage the UV has taken and whether or not the UV is waiting for an input from the operator, about each UV the operator is controlling.
- The timeline shows the estimated time it will take each UV to arrive at its next target.



**Figure 4.4: RESCHU – payload camera view (visual search task)**

For this experiment, the operators were simulated by having a subject matter expert play through RESCHU according to a script that dictated the behavior needed for each operator in each scenario. The RESCHU interfaces for each simulation were recorded via screen capture software and log files and later played back as needed to re-create each team scenario. As the team supervisor, each participant in the experiment used the RESCHU display to observe the simulated operator behavior in order to gather information about current and potential future operator performance issues.

### ***4.2.3 Team Supervisor Interface: HSMM-based Operator State Decision Support Tool***

Due to the format of the study, specifically that the subjects watched a playback of simulated scenarios and only interacted with the DST when deciding to intervene in a scenario or reset the alert, a few changes were made to the DST design explained in Chapter 3. This was done in order to provide greater experimental control by reducing the number of uncontrolled variables and degrees of interaction freedom. Thus, the DST used in the experimental evaluation was akin to a part-task trainer (Wharton, et al., 1994) that allowed the researchers to focus on the critical components which were the overall subject decision making processes and how access to the DST affected those processes.

A testbed DST screenshot is shown in Figure 4.5. The changes made to the experimental version were minor, such as not allowing the grid to be toggled on/off for the Interaction Frequency and Model Accuracy Prediction plots, not allowing the supervisor to hide portions of the Model Performance History that he/she viewed, not having UIN threshold lines present on the Interaction Frequency plot (the UIN threshold lines were set at the lower and upper bounds of the y-axis), and not having an Alert Panel for the subjects to set customizable thresholds. These changes did not impact the ability to take the measures needed for the experiment.

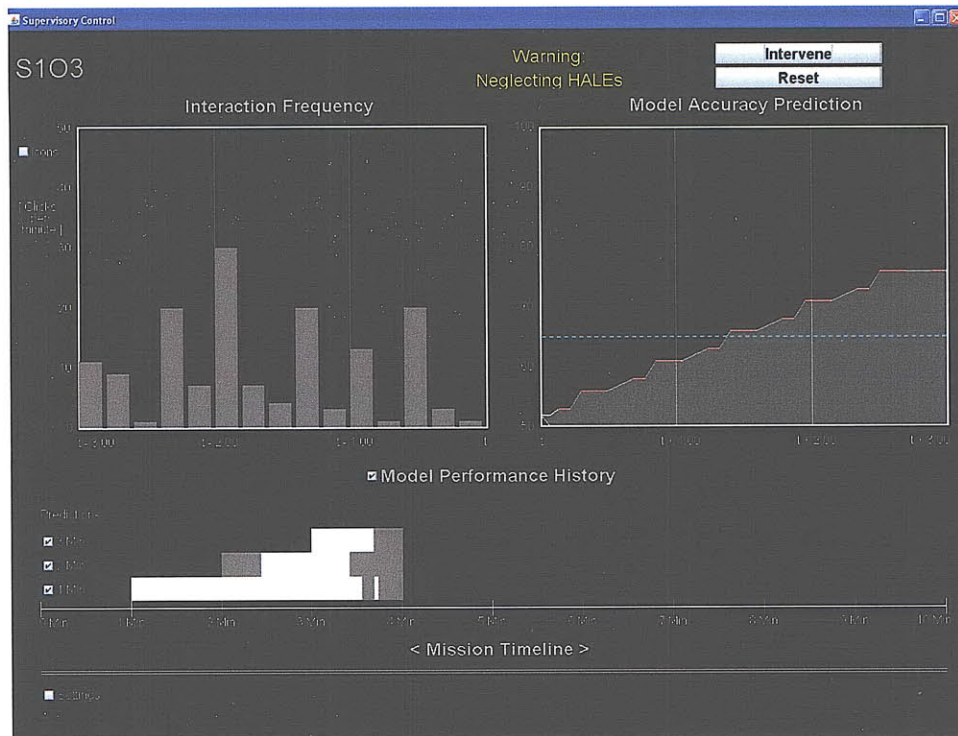


Figure 4.5: DST screenshot with alert shown

#### 4.2.4 Secondary Task Display

In order to determine if the DST had an effect on mental workload, a secondary task was used to measure the supervisor's spare mental capacity (Wickens & Hollands, 2000). The large wall screen display, shown in Figure 4.6, was used to play a different video clip for each scenario. The subjects were instructed to listen for a certain target utterance, as shown in Table 4.1, and report to the examiner the number of occurrences of that utterance at the end of the scenario. Paper and pen were provided so that the subjects would not need to remember the count. The number of missed occurrences of the target utterance was

used as an indication of the supervisor’s spare mental capacity, and thus a higher level of missed occurrences was seen as indicative of a higher level of mental workload.

**Table 4.1: Secondary Task target utterance and occurrences**

	Target Utterance	Occurrences
Scenario 1	health care	5
Scenario 2	debt	7
Scenario 3	Perot	4
Scenario 4	unpopular	6



**Figure 4.6: Large-screen wall display with scenario 1 video shown**

### **4.3 Experimental Design**

The experiment was a 2x2x2 mixed design with the following three factors:

- Operator Situation: 2 levels (true positive and false positive)
- Number of Operators Causing an Alert: 2 levels (single and multiple)
- Assistance Type: 2 levels (with DST or without DST)

The experimental design repeats measures on the Operator Situation and Number of Operators Causing an Alert factors while the Assistance Type is a between subjects factor. Descriptions of the four scenarios, as well as the randomized test matrix, are contained in Appendix C.

### ***4.3.1 Independent Variables***

#### ***Operator Situation***

This factor refers to the operator actions that cause the DST alert, specifically whether the alert requires intervention. While all alerts are caused by anomalous behavior in that the threshold value is crossed, an alert does not necessitate that the operator is performing poorly and the situation requires supervisor intervention. The alert may be caused by a sudden change in the operating environment during which the operator is correctly handling the new situation, or it may be caused by an operator that is performing in a way that the model has never experienced before and thus flags as anomalous. Since it is important that the supervisor accurately identifies whether an alert requires intervention or not, this factor is necessary. There are two factor levels for operator situation: true positive and false positive.

- *True positive* – An example of a true positive situation would be if the supervisor is alerted because the Model Accuracy Prediction dropped below the set UIN threshold, and the alert says that the operator is neglecting the underwater unmanned vehicles (UUVs). If a glance at the operator's RESCHU interface confirms that operator is neglecting the UUVs, then this type of situation requires supervisor intervention.
- *False positive* – An example of a false positive situation would be if the supervisor is alerted that an operator is operating below the low interaction frequency threshold. If a glance at the operator's RESCHU interface confirms that all of the UVs are on path toward targets that are far away from their current positions, and the operator does not have anything else to do at that time, then this type of situation does not require supervisor intervention. While this situation is anomalous in that operators rarely interact at such a low frequency, resulting in a drop below the threshold value, the operator is not doing anything wrong.

#### ***Number of Operators Causing an Alert***

The number of operators causing an alert is broken into two levels: single and multiple. This factor corresponds to the number of simultaneous alerts that the supervisor receives due to the number of operators that are exhibiting anomalous behavior. In the multiple alert scenarios, one alert is caused by true positive behavior, while the other alert is caused by false positive or true positive behavior based on the Operator Situation factor level. It was hypothesized that the multiple alert scenarios would result in slower response times and a higher secondary task ratio than the single alert scenarios due to the simultaneous nature of the alerts.



### *Assistance Type*

The two levels of assistance type are with the DST and without the DST. It was hypothesized that subjects who used the DST would perform better on all dependent variables than non-DST users.

### **4.3.2 Dependent Variables**

#### *Decision Accuracy*

Decision Accuracy is a binary yes/no value of whether the subject correctly identified the need to intervene or not intervene for each operator situation that caused an alert.

#### *Incorrect Interventions*

An Incorrect Intervention was recorded if a subject chose to intervene in a situation when the operator was acting normally and an intervention was not necessary. This metric does not include decisions made in response to the operator actions that caused an alert. For example, if the correct response to the operator actions that caused an alert was to not intervene and the subject chose to intervene, this was considered an incorrect decision and affected Decision Accuracy. This example would not be counted as an Incorrect Intervention. A choice to intervene during normal operator actions that did not correspond to one of the alerts was recorded as an Incorrect Intervention.

#### *Response Time*

Response Time is the amount of time in seconds that passes from when an alert is triggered until the supervisor makes the decision of whether to intervene or acknowledge/reset the alert. For non-DST users, the Response Time is calculated from the mission time that the alert would have been triggered if they were using the DST.

#### *Secondary Task Ratio*

The Secondary Task Ratio is the difference between the reported total number of occurrences of the target utterance in the secondary task video and the actual total number of occurrences of the target utterance in the video, normalized by the actual total number of occurrences.

$$\text{Secondary Task Ratio} = \frac{|\# \text{ Reported Occurrences} - \# \text{ Actual Occurrences}|}{\# \text{ Actual Occurrences}}$$

### ***DST Understanding***

A verbal retrospective protocol was used in conjunction with an analysis of the subjects' interactions with the DST in order to infer higher level strategies and understanding. Additionally, the Post Experiment Questionnaire in Appendix D was used to obtain subjective feedback from subjects that used the DST.

### ***4.4 Procedure***

After being welcomed and thanked for participating in the experiment, each subject read and signed a consent form (Appendix E) and filled out a demographic survey (Appendix F). The experimental task was then explained in detail through a self-paced presentation of the applicable training slides (Appendices G and H, respectively). The subject was given a chance to ask any questions before moving into the first of two practice trials. The practice trials and each of the four experimental trials were all controlled from a networked computer. The experimenter talked through the two practice trials with the subject in order to point out and reinforce important information from the training slides. In the second half of the second practice trial, the experimenter stopped talking and allowed the subject to continue the scenario as if it were an experimental trial.

Next, the subject was given another chance to ask any questions before moving on to the four experimental trials, one for each combination of the Operator Situation and Number of Operators Causing an Alert factor levels as shown in Table 4.2. Assistance Type was a between subjects factor. These trials lasted approximately five minutes each and were presented in the counterbalanced order shown in Appendix C. During each trial, the experimenter took notes on subject behavior such as scan pattern, how each subject kept track of the secondary task occurrences, the mission time at which each decision to intervene or reset an alert was made, and anything else of notable interest.

**Table 4.2: Experimental scenario descriptions**

	<b>Scenario Description</b>	<b>Operator 1 Behavior</b>	<b>Operator 2 Behavior</b>	<b>Operator 3 Behavior</b>
<b>Scenario 1</b>	true positive single alert	normal	normal	True alert - Neglect HALEs (t=3:30)
<b>Scenario 2</b>	true positive multiple alerts	True alert - low interaction frequency (t=3:20)	True alert - excessive status checking (t=3:20)	normal
<b>Scenario 3</b>	false positive multiple alerts	normal	False alert - Neglect HALEs (t=2:15)	True alert - excessive waypoint modification (t=2:15)
<b>Scenario 4</b>	false positive single alert	False alert - low interaction frequency (t=1:14)	normal	normal

After each trial, the experimenter re-started the same scenario, this time without the secondary task video, and conducted a verbal retrospective protocol during which the subjects explained their thought process throughout the scenario. The experimenter also asked questions in an open-ended interview format in order to gain more insight, such as what portions of the displays most helped the subject in making the crucial decision of whether to intervene or not. After the experimental trials were completed, DST subjects filled out the post-experiment questionnaire in Appendix D. All subjects were then paid a nominal fee and thanked for their participation. In order to encourage subjects to perform their best, they were instructed before the experiment began that the top performer, based on the dependent variables listed in Section 4.3.2, would receive a \$200 gift card.

#### ***4.5 Experimental Evaluation Summary***

As with any design, it is important to test prototypes early in the design cycle in order to ensure the design warrants moving forward to more expensive production and testing methods. This chapter explained the experimental evaluation designed to evaluate the proposed DST design from Chapter 3. Each subject participated in four different trials as the team supervisor of three simulated operators in order to provide data to be used to evaluate the effectiveness of the DST on Decision Accuracy, the number of Incorrect Interventions, Response Time, and the Secondary Task Ratio. Independent variables included Operator Situation, Number of Operators Causing an Alert, and Assistance Type, making this experiment a 2x2x2 study. The next chapter contains the results from this experiment.



## 5. Results and Discussion

This chapter presents the results of the experiment described in Chapter 4. The descriptions of each of the four experimental scenarios, from Table 4.2, are reproduced below for quick reference. Recall that the multiple alerts in Scenarios 2 and 3 were simultaneous; they appeared at the exact same time for DST users.

**Table 4.2: Experimental scenario descriptions**

	<b>Scenario Description</b>	<b>Operator 1 Behavior</b>	<b>Operator 2 Behavior</b>	<b>Operator 3 Behavior</b>
<b>Scenario 1</b>	true positive single alert	normal	normal	True alert - Neglect HALEs (t=3:30)
<b>Scenario 2</b>	true positive multiple alerts	True alert - low interaction frequency (t=3:20)	True alert - excessive status checking (t=3:20)	normal
<b>Scenario 3</b>	false positive multiple alerts	normal	False alert - Neglect HALEs (t=2:15)	True alert - excessive waypoint modification (t=2:15)
<b>Scenario 4</b>	false positive single alert	False alert - low interaction frequency (t=1:14)	normal	normal

All dependent variables were analyzed using non-parametric tests because the parametric assumptions of normality and/or homogeneity of variance were not met. Furthermore, an  $\alpha$  value of 0.10 was used for all statistical tests. Since subjects had relatively little time to become familiar with the system, Kruskal-Wallis tests were performed on the order the scenarios were presented. None of the dependent variables showed a statistically significant learning effect. These tests, as well as all supporting statistics from this chapter, are included in Appendix I. The results for each dependent variable are described below.

### 5.1 Decision Accuracy

Decision Accuracy was analyzed by comparing the correct response to each alert with the decision each subject made. The correct response, either to intervene in the scenario or determine that intervention was not necessary and reset the alert, was determined a priori via the scripted behavior of the different operators in each scenario. The number of correct decisions for DST users was then compared with the number of correct decisions for non-DST users with Mann-Whitney U tests.<sup>3</sup> For the single alert scenarios, the possible number of correct decisions was zero or one per scenario. For the multiple alert scenarios, the possible number of correct decisions was zero, one, or two per scenario. The Decision

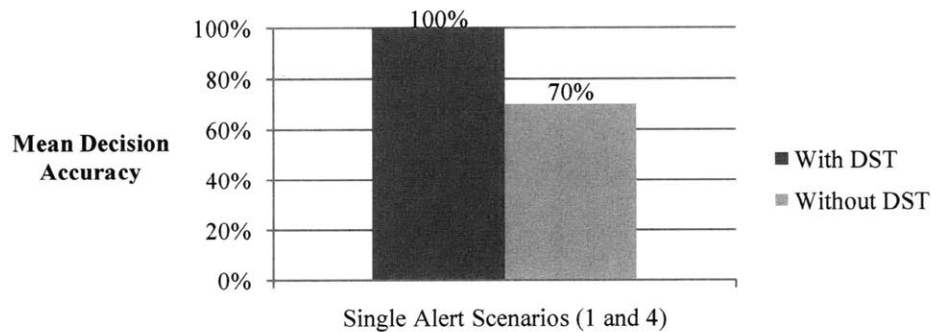
<sup>3</sup>The Mann-Whitney U test is a non-parametric test used to determine if two independent samples of observations are from the same distribution.

Accuracy results for DST and non-DST users are shown in Table 5.1 (statistically significant results are highlighted in gray).

**Table 5.1: Decision Accuracy results**

	Decision Accuracy	
	With DST	Without DST
Scenario 1	100%	47%
Scenario 2	73%	83%
Scenario 3	77%	87%
Scenario 4	100%	93%
True Positive Scenarios (1 and 2)	82%	71%
False Positive Scenarios (3 and 4)	84%	89%
Single Alert Scenarios (1 and 4)	100%	70%
Multiple Alert Scenarios (2 and 3)	75%	85%
All Scenarios	83%	80%

The scenarios were analyzed individually as well as by each independent variable. In Scenario 1, DST users were shown to have a significantly higher Decision Accuracy than non-DST users ( $U = 52.5$ ,  $n_1 = n_2 = 15$ ,  $p = 0.001$ ). When analyzing the Decision Accuracy of the single alert scenarios together (Scenarios 1 and 4), it can be seen in Figure 5.1 that non-DST users made the correct decision 70% of the time (21 of 30), as compared to DST users who made the correct decision 100% of the time (30 of 30). A Mann-Whitney U test proved that this was a statistically significant result ( $U = 315.0$ ,  $n_1 = n_2 = 30$ ,  $p = 0.001$ ).



**Figure 5.1: Mean Decision Accuracy per subject**

Nine Mann-Whitney U tests were run on this dependent variable, thus Kimball's Inequality<sup>4</sup> requires individual tests to have  $p < 0.012$  in order to maintain a family-wise  $\alpha$  value of 0.10. The two tests

<sup>4</sup>  $\alpha_{family-wise} = 1 - (1 - \alpha_{each\ test})^{\# tests}$

reported as statistically significant meet this stricter standard. So, while the DST was not shown to have an effect on Decision Accuracy for subjects that responded to two simultaneous alerts (Scenarios 2 and 3 grouped together), the DST did have a positive effect on Decision Accuracy for subjects that responded to one alert at a time (Scenarios 1 and 4 grouped together).

## 5.2 *Incorrect Interventions*

An Incorrect Intervention was recorded if a subject chose to intervene in a situation when the operator was acting normally and an intervention was not necessary. This does not include decisions made in response to operator actions that caused an alert. For example, if the correct response to operator actions that caused an alert was to not intervene and the subject chose to intervene, this was considered an incorrect decision and affected Decision Accuracy. This example would not be counted as an Incorrect Intervention. In contrast, a choice to intervene during normal operator actions that did not correspond to one of the alerts was recorded as an Incorrect Intervention. The mean number of Incorrect Interventions made by subjects with respect to the scenario type and whether or not they were using the DST are shown in Table 5.2 (statistically significant results are highlighted in gray).

**Table 5.2: Incorrect Interventions results**

	Mean Number of Incorrect Interventions per Subject	
	With DST	Without DST
Scenario 1	0.40	1.40
Scenario 2	0.13	0.67
Scenario 3	0.07	0.73
Scenario 4	0.00	0.20
True Positive Scenarios (1 and 2)	0.27	1.03
False Positive Scenarios (3 and 4)	0.03	0.47
Single Alert Scenarios (1 and 4)	0.20	0.80
Multiple Alert Scenarios (2 and 3)	0.10	0.70
All Scenarios	0.15	0.75

Non-DST users made significantly more Incorrect Interventions than DST users in all scenarios ( $U = 1174.5$ ,  $n_1 = n_2 = 60$ ,  $p < 0.001$ ). This result was consistent throughout individual scenario analysis (Figure 5.2) and analysis by independent variable grouping. All relevant statistics are provided in Appendix I. Nine Mann-Whitney U tests were run on this dependent variable, thus Kimball's Inequality requires individual tests to have  $p < 0.012$  in order to maintain a family-wise  $\alpha$  value of 0.10. The test on all scenarios meets this stricter standard.

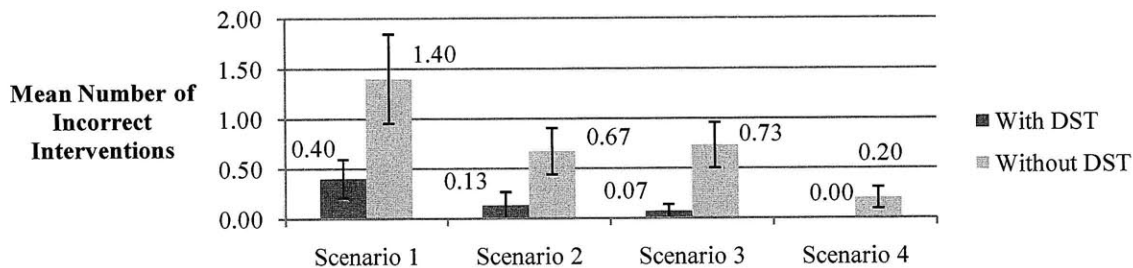


Figure 5.2: Mean Number of Incorrect Interventions per subject

### 5.3 Response Time

Response Time was analyzed by recording the amount of time in seconds that passed from when operator actions caused an alert, either true positive or false positive, until the supervisor made the decision of whether to intervene. For DST users, the time was recorded when the subject clicked either the Intervene or Acknowledge/Reset Alert button. For non-DST users, the time was recorded when the subject notified the experimenter that they wanted to intervene by saying the word “Intervene” as instructed in the pre-experiment training. For non-DST users, this “Intervention” time was then compared to the mission time that an alert would have occurred if the subject had been using the DST since the non-DST users did not receive any alerts. There was no way to record when the non-DST users made decisions to “Acknowledge/Reset Alert” since they did not receive alerts, and thus only “Intervene” Response Times were recorded for the non-DST users.

Additionally, since both the non-DST users and the DST users could intervene whenever they wanted, it was possible for subjects to recognize a problem requiring intervention before the mission time when an alert would occur. If the supervisor did intervene before the mission time when the alert would occur, this resulted in a negative Response Time and is referred to as an Early Intervention. Thus, negative Response Times/Early Interventions were possible and are shown in the results. Early Interventions were most prevalent in Scenario 2, when an operator abruptly stopped interacting with the RESCHU interface. Many supervisors, especially non-DST users, recognized this problem in just a few seconds, while the algorithm required a couple state transitions before identifying the deficiency and thus took several seconds before presenting an alert to the supervisor.

The mean first Response Times per subject (the response to the first alert in each scenario) are shown by individual scenario in Figure 5.3 and by relevant groupings in Figure 5.4. Non-DST users were shown to have a quicker first Response Time than DST users when analyzing all four scenarios together ( $U =$



742.0,  $n_1 = 35$ ,  $n_2 = 60$ ,  $p = 0.017$ ). However, the scenario that allowed the most direct comparison, the true positive, single alert scenario (Scenario 1), showed a statistically significant result that the non-DST users had a slower response time than the DST users ( $U = 26.5$ ,  $n_1 = 7$ ,  $n_2 = 15$ ,  $p = 0.067$ ). This scenario was analyzed separately, as opposed to with the false positive, single alert scenario (Scenario 4), because the Response Time data for non-DST users was not obtainable if they did not intervene in the scenario, and Scenario 4 did not require an intervention. Thus, there was only one data point for a non-DST user first Response Time in Scenario 4. The multiple alert grouping (Scenarios 2 and 3 together) resulted in faster mean Response Times for non-DST users as compared to DST users ( $U = 241.5$ ,  $n_1 = 27$ ,  $n_2 = 30$ ,  $p = 0.009$ ). Non-DST users also had faster mean Response Times than DST users for Scenario 2 ( $U = 40.0$ ,  $n_1 = n_2 = 15$ ,  $p = 0.003$ ). Six Mann-Whitney U tests were run on this dependent variable, thus Kimball's Inequality requires individual tests to have  $p < 0.0174$  in order to maintain a family-wise  $\alpha$  value of 0.10. The tests run on Scenario 2, the multiple alert grouping (Scenarios 2 and 3 together), and all scenarios grouped together meet this stricter standard.

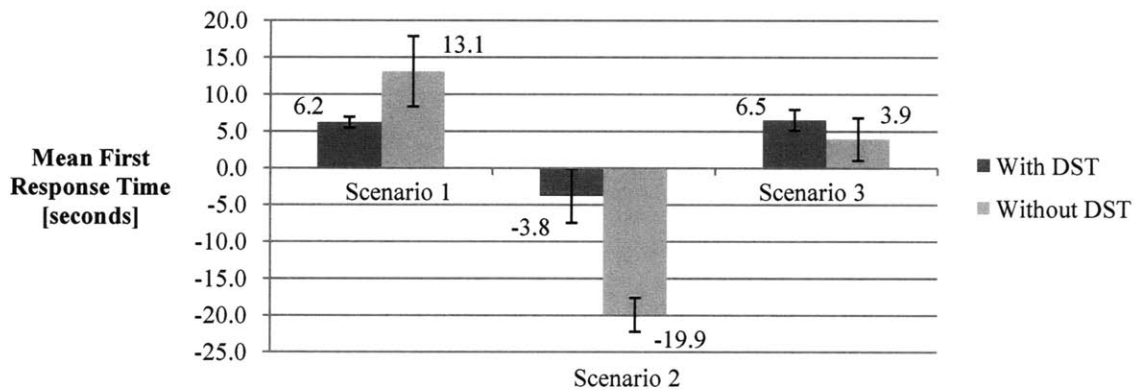


Figure 5.3: Mean first Response Times per subject (individual scenarios)

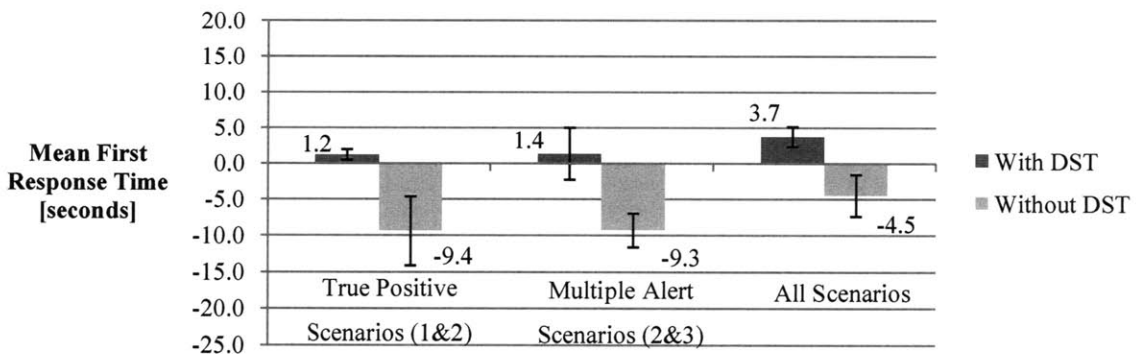
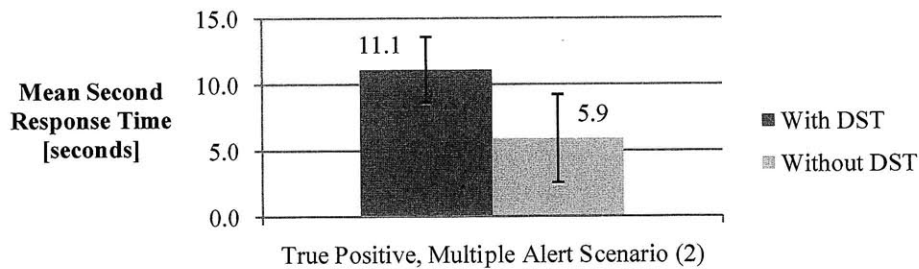


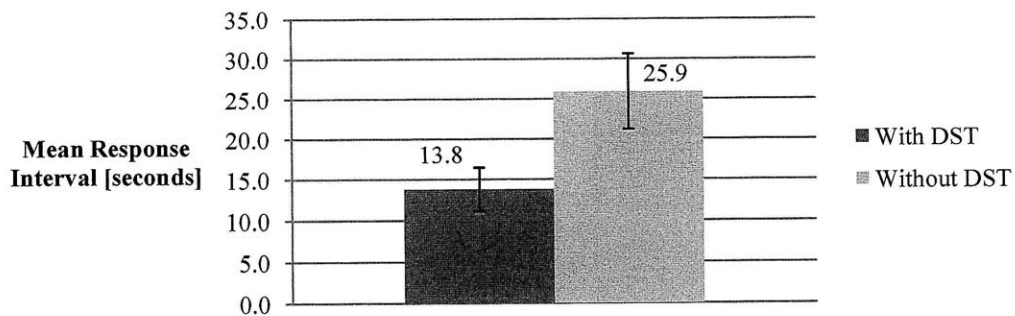
Figure 5.4: Mean first Response Times per subject (grouped scenarios)

Analysis of the second Response Time was limited to the multiple alert scenarios, since they were the only scenarios that required two responses. The second Response Time can also be considered the total Response Time, since it is the total amount of time, in reference to the alerting time, that each subject took in order to respond to both operator situations. The second Response Time data for non-DST users was not obtainable if they did not intervene in the scenario. Thus, the false positive, multiple alert scenario (Scenario 3), which only required one intervention, resulted in only one data point for a non-DST user second Response Time. Therefore, since the true positive, multiple alert scenario (Scenario 2) required two responses, it was analyzed separately (Figure 5.5). There was not a statistically significant difference in second Response Time (total Response Time) between non-DST and DST users for the true positive, multiple alert scenario ( $U = 45.0, n_1 = 10, n_2 = 12, p = 0.322$ ).



**Figure 5.5: Mean second Response Times (total Response Time) per subject**

However, non-DST users did have a significantly longer interval of time between first responses and second responses for the true positive, multiple alert scenario (Scenario 2) when compared to DST users ( $U = 28.5, n_1 = 10, n_2 = 12, p = 0.038$ ). This can be seen in Figure 5.6.



**Figure 5.6: Mean response interval per subject**

An Early Intervention corresponds to a situation where the subject intervenes in the scenario before the alert occurs for a DST user, or before the mission time when the alert would have occurred for a non-DST user. The number of Early Interventions for each scenario is shown in Table 5.3. A majority of the Early

Interventions took place during Scenario 2, which included the operator who abruptly stopped moving his cursor on the screen.

**Table 5.3: Early Interventions**

	Early Interventions	
	With DST	Without DST
Scenario 1	0	1
Scenario 2	7	18
Scenario 3	1	3
Scenario 4	0	0
All Scenarios	8	22

In summary, the DST was shown to have a positive effect on Response Time for the true positive, single alert scenario (Scenario 1), a negative effect on first Response Time in the multiple alert scenarios (Scenarios 2 and 3), and no effect on total Response Time. Additionally, non-DST users appeared to intervene before the alert time more often than DST users, and non-DST users took a longer amount of time between responding to the two positive alerts (Scenario 2).

#### **5.4 Secondary Task Ratio**

The Secondary Task Ratio is defined as the difference between the reported total number of occurrences of the target utterance (such as the word “debt”) in the secondary task video and the actual total number of occurrences of the target utterance in the video, normalized by the actual total number of occurrences. For example, a Secondary Task Ratio of 0.0 would indicate that the operator reported the correct number of utterances. The Secondary Task Ratio results for DST and non-DST users are shown in Table 5.4. There was not a significant difference between non-DST users and DST users for any of the scenarios or groupings of scenarios, which indicates that the DST did not have a significant impact on secondary task workload.

**Table 5.4: Secondary Task Ratio results**

	Mean Secondary Task Ratio per Subject	
	With DST	Without DST
Scenario 1	0.37	0.40
Scenario 2	0.34	0.37
Scenario 3	0.35	0.35
Scenario 4	0.25	0.21
All Scenarios	0.33	0.33

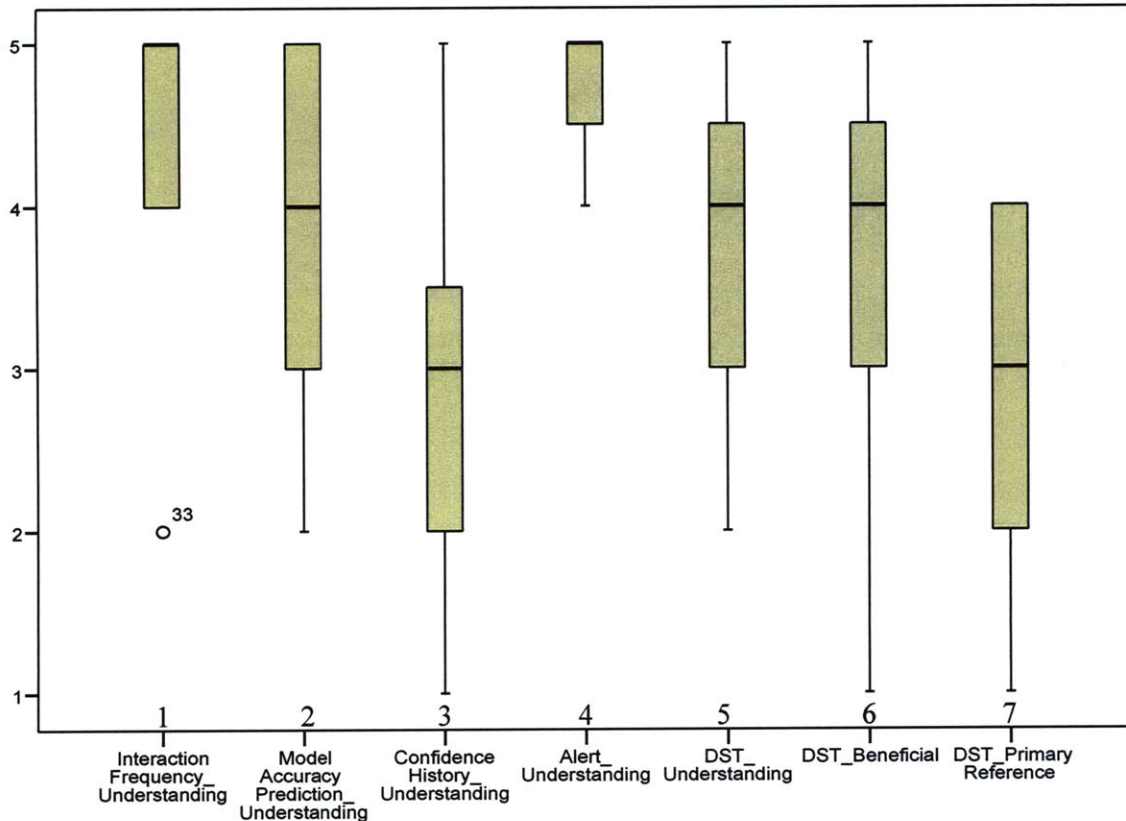
### 5.5 Subjective DST Understanding

DST users answered the first seven questions in the post-experiment questionnaire in Appendix D, reproduced in Table 5.5, using the following Likert scale:

- 1 – Strongly disagree
- 2 – Somewhat disagree
- 3 – Neutral
- 4 – Somewhat agree
- 5 – Strongly agree

**Table 5.5: DST Understanding - Likert scale descriptive statistics**

	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
1. It was easy for me to understand the data presented in the Interaction Frequency plot.	2	5	4.4	5	5	.83
2. It was easy for me to understand the data presented in the Model Accuracy Prediction plot.	2	5	3.8	4	4,5	1.15
3. It was easy for me to understand the data presented in the Confidence History plot.	1	5	2.9	3	3	1.03
4. It was easy for me to understand the DST alerts.	4	5	4.7	5	5	.46
5. It was easy for me to understand the data presented in the DST as a whole.	2	5	3.8	4	4	1.01
6. The information presented on the DST was beneficial in helping me make my decision of whether to intervene or not intervene.	1	5	3.6	4	4	1.35
7. I primarily referenced the DST when making my decision to intervene or not intervene.	1	4	2.9	3	4	1.19



**Figure 5.7: DST Understanding - Likert scale box plots**

As can be seen by the results shown in Table 5.5 and Figure 5.7, user understanding of the Interaction Frequency plot (#1) and DST alerts (#4) was quite high. User understanding of the Model Accuracy Prediction plot (#2) had an above neutral median with a wide range of responses. User understanding of the Confidence History plot (#3) had a neutral median. Users seemed to agree that they were able to easily understand the data presented in the DST as a whole (#5) and that the information was beneficial in helping to make the decision of whether to intervene or not intervene (#6). However, users did not seem to primarily reference the DST when making the decision of whether to intervene (#7). This finding was reinforced by many comments from DST users (8 subjects out of 15) stating that the DST was useful in alerting them to scenarios they might not have otherwise seen and gave them a good idea of what to look for in the scenario, but users gathered detailed information about the situation primarily from the RESCHU interface in order to decide whether to intervene. Overall, the DST received neutral to positive feedback (medians 3-5) overall all seven questions that were asked.

## ***5.6 Discussion of Experimental Findings***

Summarizing the results, the proposed DST has been shown in single alert scenarios to improve performance in terms of increased Decision Accuracy, decreased Incorrect Interventions, and decreased Response Time (for the single alert, true positive scenario only), while having no effect on supervisor mental workload (as measured by a secondary task). These results were obtained for a team supervisor of three simulated operators, each independently controlling a set of highly autonomous UVs. In scenarios where the supervisor faced multiple, simultaneous alerts, the DST was shown to decrease the number of Incorrect Interventions, and have no effect on Decision Accuracy, total Response Time, or supervisor mental workload (as measured by a secondary task). Subjects reported high understanding of the Interaction Frequency plot and DST alerts while expressing difficulty with the Confidence History plot. Overall, users found the information in the DST beneficial in making the decision of whether to intervene, but did not primarily reference the DST when making the decision of whether to intervene. These results are discussed below.

### ***5.6.1 Decision Accuracy***

The DST gave subjects a possible cause of each alert, such as “low interaction frequency,” and thus gave DST users an indication of what type of information to analyze when making the decision of whether to intervene. This focused the attention of DST users to the likely cause of the alert, and whether the alert required action, as opposed to the non-DST users that had to continuously search through all available information to determine if a situation required intervention. Subjects were told their performance would be evaluated first on their ability to correctly decide when to intervene (maximizing Decision Accuracy and minimizing Incorrect Interventions) and second on Response Time. Therefore, the subjects knew they were in a time critical situation and needed to make decisions quickly. This added pressure of needing to make a decision quickly likely led to subjects that made decisions without full confidence that they were correct. For DST users, having their attention focused on the mostly likely cause of the alert may have been the reason for improved Decision Accuracy in the single alert scenarios, both the true positive and false positive variants.

While the DST was not shown to have a significant impact on Decision Accuracy for multiple alert scenarios, this result is possibly due to the simultaneous nature of the alerts. The DST users may have felt that they needed to rush their first response after being alerted in order to quickly work on the second alert, and thus sacrificed accuracy for time. Non-DST users did not have this same dynamic as there were no alerts, and thus the non-DST users could finish dealing with the first operator who seemed to be acting

abnormally without knowing that another operator was in a situation waiting to be addressed. The non-DST users serially dealt with one alert at a time, while the DST users dealt with the parallel recognition of two simultaneous alerts. Although it is conceivable that non-DST users could identify two problems in parallel as well, this does not seem to be the case. Non-DST users had a significantly longer interval of time between first responses and second responses in the multiple alert, true positive scenario (Scenario 2) when compared to DST users.

This difference between parallel and serial recognition of alerts may have influenced why the DST did not impact Decision Accuracy in the multiple alert scenarios; humans can only solve one complex problem at a time (Broadbent, 1958; Nehme, 2009; Welford, 1952). Therefore, the DST users' focused attention advantage from the single alert scenarios may have been counteracted by the negative influence of having to simultaneously react to two alerts. A future study should evaluate multiple alert scenarios that have different spacing between alerts in order to determine if the DST has an impact in non-simultaneous multiple alert scenarios.

### ***5.6.2 Incorrect Interventions***

Frequent Incorrect Interventions by a supervisor may have negative consequences on team performance because operators may become irritated, experience a loss of situational awareness, or lose trust in the team supervisor (Brewer & Ridgway, 1998; Irving, et al., 1986; Mitchell, et al., 2004). When analyzing the effect of the DST on the number of Incorrect Interventions, it is important to note that the DST notified the subjects of possible abnormal operator behavior, while non-DST users had no such help. DST users were also given trend information in the Model Accuracy Prediction plot that allowed the DST users to predict when alerts might occur. This allowed subjects to identify and analyze suspicious operator behavior before being alerted. Many of the DST users reported that they assumed operator behavior was normal if they were not alerted to a situation and trusted they would be alerted when abnormal behavior was occurring. This automation bias (Cummings, 2004) may result in substantial misses of abnormal behavior that the system does not provide an alert for (i.e., Type II error) and should be studied in future work.

Despite the training scenarios that attempted to help non-DST users identify abnormal behavior without the aid of the DST, and thus understand how abnormal was “abnormal enough” to intervene, the non-DST users may not have fully understood what type of operator actions required intervention. Many subjects described strategies such as looking for blinking lights (blinking lights were indicative of UVs that were waiting for operator input, such as to begin a visual task or engage a target) and making sure the operators

were moving their input devices and interacting with the interface. Very few subjects described more complex strategies such as ensuring the operators were prioritizing correctly with the different types of UVs and tasks. This lack of understanding of operator actions, combined with the absence of alerts, may have been the reason non-DST users had more Incorrect Interventions. However, there was not a statistically significant learning effect for any of the dependent variables (Table I.22), suggesting that additional training may not be enough to decrease the number of non-DST user Incorrect Interventions. These same results may not be found with subjects who have experience with the underlying operator control system, RESCHU in the case of this experiment, and this factor should be evaluated in future work.

### ***5.6.3 Response Time***

The overall slower first Response Time (all scenarios grouped together) for DST users may have been caused by the fact that DST users had more information to process via the additional plots and text on the DST. Additionally, the effect may have been caused by the same automation bias that resulted in DST users having a lower number of Incorrect Interventions; DST users were unlikely to intervene without being alerted. Some DST users even stood at the DST interface waiting to click the Intervene or Acknowledge/Reset Alert button because they deduced an alert was coming based on the trend of the Model Accuracy Prediction plot, but waited until that alert occurred to take action. Only 8 instances occurred of DST users intervening in a situation before an alert, as opposed to 22 instances of non-DST users intervening in a situation before the alert would have occurred had they been using the DST. The additional information that DST users had to process, combined with their bias toward not intervening before being alerted, may be the reason DST users had overall slower first Response Times than non-DST users.

It is worth noting that in operational settings, the timing of each alert is affected by the user-set threshold values. Thus, giving the user the option to change the threshold values could result in dramatically different response times. This option was included in the design of the interface but was not made functional in this experiment in order to decrease variance between subjects and provide more experimental control. The analysis of this feature should be studied in future work.

DST users did, however, have quicker Response Times in the single alert, true positive scenario. This quicker response may have been caused by the serial recognition of the single alert as opposed to the parallel recognition of the multiple alert scenarios, similar to Decision Accuracy. In multiple alert scenarios, DST users could not focus solely on an individual alert, since they had to choose one of the



simultaneous alerts to respond to first before moving on to the second alert. While this could greatly increase the second Response Time (total Response Time) for DST users as they had to first take care of one alert before reorienting to the other operator to deal with the second alert, this was not shown to be the case. DST users and non-DST users were not shown to have differing second Response Times (total Response Time).

Additionally, since each alert was for a different type of operator behavior, it is possible that some types of abnormal behavior are easier for humans to identify before being alerted and thus certain types of operator behavior may result in increased early interventions. This seemed to be the case in Scenario 2 when one of the operators suddenly stopped interacting with the interface. A low interaction frequency alarm was triggered, but not for several seconds after the operator's cursor stopped moving. The lack of cursor movement was quickly recognized by the non-DST subjects and resulted in many Early Interventions for Scenario 2. This quick recognition would become more difficult for a supervisor of more than 3 operators, and thus the scalability of the DST system should be evaluated in future work.

#### ***5.6.4 Secondary Task Ratio***

The DST was not shown to have a significant impact on mental workload as measured by the secondary task. Since the addition of the DST to the RESCHU interface doubled the number of screens, the team supervisor needed to monitor from 3 to 6, the DST would have theoretically increased mental workload if it were poorly designed. The increased Decision Accuracy, decreased Incorrect Interventions, and quicker Response Time for DST users in single alert scenarios did not come at the cost of increased mental workload as shown by the statistically insignificant results.

Additionally, there was anecdotal evidence that the DST may help to lower mental workload if the supervisor has a team of more than 3 operators. All 8 of the DST users who reported relying primarily on the RESCHU interface for the intervention decision stated they would rely on the DST more if the number of operators was increased. Therefore, the scalability of the RESCHU-DST system should be tested in future work to see if the RESCHU-DST system can still maintain or even decrease mental workload of a supervisor in charge of a larger number of operators.

### ***5.6.5 Subjective Feedback***

Subjective feedback showed that user understanding of the Interaction Frequency plot and DST alerts was very high. This high understanding is not surprising as these were very simple aspects of the DST that did not use any information from the HSMM-based operator model.

User understanding of the Model Accuracy Prediction plot varied. Of the 15 DST users, 9 reported that the Model Accuracy Prediction plot was the most beneficial aspect of the DST in deciding whether to intervene. This feedback indicates that the Model Accuracy Prediction plot was useful, but one user reported that he would make the trend lines larger in order to facilitate quicker Response Times. Another user stated that the Model Accuracy Prediction plot did not seem useful outside of the window between current mission time and +1:00 minute. This forecasting problem of determining how far in the future the predictions from the model are beneficial for the user could be evaluated in future work.

Overall, the Confidence History plot seems to require a redesign, if not deleted altogether. It received the lowest feedback for user understanding, and 7 of 15 DST users made specific comments about how they did not understand or did not use the Confidence History plot. There are many possible causes for this response. One possible cause could be that the training for this study did not clearly explain the Confidence History plot or focused too much on the math that occurs in the background in order to produce the Confidence History plot. Additionally, the concept of comparing past performance on different prediction horizons was difficult for subjects to grasp. The design of the Confidence History plot should be re-evaluated to determine if it is beneficial to keep in the DST, and if so, what changes should be made in order to ensure its usefulness.

As stated in Section 5.5, the DST was useful in alerting users to scenarios they might not have otherwise seen and gave them a good idea of what to look for in the scenario, but users gathered detailed information about the situation primarily from the RESCHU interface in order to decide whether to intervene. It may be useful to see how supervisors equipped with only the DST alerts, or possibly the DST alerts and interaction frequency plot, would perform as compared to supervisors equipped with the DST in its entirety. Perhaps the Model Accuracy Prediction plot and Confidence History plot do not significantly influence the overall effectiveness of the DST. One subject reported, “It was hard trying to read three graphs during a time-sensitive, stressful situation. Maybe digital readings would be more helpful...” Future work will be needed to evaluate this claim.

### ***5.6.6 Summary of Results***

In summary, the proposed DST was shown to increase Decision Accuracy, decrease Incorrect Interventions, and decrease Response Times (single alert, true positive scenario only) in single alert scenarios, while not affecting supervisor mental workload. However, the current design does not produce the same positive results in multiple alert scenarios. There are several possible reasons for this variance, including difficulties associated with parallel recognition of simultaneous alerts and DST users exhibiting possible automation bias in not intervening without an alert. This automation bias could prove detrimental to overall system performance if the underlying algorithm, combined with user settings, results in a situation where the supervisor is not alerted to an operator problem. Additionally, the quantitative and qualitative results obtained from this experiment have highlighted areas where design changes to the DST may result in performance increases across multiple alert domains. More importantly, the experiment showed that recent advances in HSMM-based operator modeling techniques can be leveraged in a real-time decision support tool to improve team supervisor performance. In the next chapter, conclusions and future work are addressed.



## 6. Conclusions

The goal of this thesis was to develop and evaluate a real-time decision support tool to determine whether team supervisors with the HSMM-based DST perform better than supervisors without the DST in terms of correctly identifying current and possible future problems, solving current problems efficiently and accurately, correctly preventing future problems from occurring, and limiting unnecessary interventions. This thesis also sought to address the question of how to best display complex probabilistic data, specifically the HSMM-based operator model output, so that someone with little to no background in statistical inferential reasoning can efficiently and accurately make time-critical decisions. These research questions were addressed through the following methods:

- A CTA was conducted in order to understand the decisions the supervisor would need to make in this domain and to determine what information the supervisor would need in order to quickly and accurately make those decisions. Design requirements were then derived from the CTA in order to focus the design cycle (Chapter 3).
- The design requirements derived from the CTA were used to design the DST. The DST takes into account the applicable literature in Chapter 2, and a discussion of the design of the DST is included in Chapter 3.
- A human subject experiment was conducted in a simulated team supervisory setting in which subject performance while using the DST was compared to subject performance in a standard team HSC setting without the DST. Details about the design of this experiment are included in Chapter 4. The experimental results and discussion are included in Chapter 5.

The answers to the research questions will now be addressed, followed by future work.

### 6.1 *Supervisor Performance*

The DST was shown to increase supervisor Decision Accuracy, decrease Incorrect Interventions, and decrease Response Times (single alert, true positive scenario only) in single alert scenarios while not affecting supervisor mental workload. The current design does not, however, produce the same positive results in multiple alert scenarios. DST users in multiple alert scenarios had significantly less Incorrect Interventions, but demonstrated no difference in Decision Accuracy or total Response Time, and had slower first Response Times than the non-DST users. There are several possible reasons for this discrepancy between single and multiple alert scenarios. One reason may be due to the DST users having more information to process for two simultaneous alerts than non-DST users because of the addition of the DST. Another reason may be due to DST users showing an automation bias in not intervening without an alert, which led to slower first Response Times in the multiple alert scenarios, especially Scenario 2.

Subjective feedback suggested that simple design changes such as separating the Intervene and Acknowledge/Reset Alert buttons may result in quicker first Response Times as users would not need to be as precise with mouse position when making sure to not accidentally click the wrong button.

## **6.2 *Display of Probabilistic Data***

Subjects showed mixed feelings toward the way that the HSMM-based operator modeling output was displayed. The Model Accuracy Prediction plot was well understood (median response = 4 of 5). Many users commented on how they relied on the future trend data in order to decide which operators they should be paying more attention to, thus leading to increased attentiveness and situational awareness when operators began to show abnormal behavior. This method of displaying probabilistic data with trend lines and uncertainty visualized as background shading seemed beneficial in this application.

Alternatively, the Confidence History plot was not well understood (median response = 3 = neutral). Almost half of the users (7 of 15) reported the Confidence History plot was confusing and not very helpful. This representation of past model performance data, multiple color-coded horizontal bar graphs, was not beneficial. It was especially difficult for users to understand the difference between the three horizontal bars as applied to the different predictions (1 minute, 2 minute, and 3 minute) and what exactly the colors meant in relation to past model performance. These results suggest the Confidence History plot needs to be redesigned in order to more intuitively and effectively communicate the historical performance of the model. This issue of the Confidence History plot provides another example in the challenge of displaying probabilistic information to humans that are often biased and subjective when given even simple probabilistic values (Tversky & Kahneman, 1974).

## **6.3 *Future Work***

While this work has shown that a real-time DST that leverages HSMM-based operator modeling techniques increased team supervisor performance in terms of increased Decision Accuracy, decreased Incorrect Interventions, and decreased Response Times (single alert, true positive scenario only) in single alert scenarios, the DST is still a prototype and not ready for operational implementation. The experiment that was conducted had limitations, such as the subjects having less than 15 minutes of training on the system before being evaluated, the subjects knowing that the operators were simulated, the inability for the subjects to control many features of the DST such as changing the UIN levels, the independence of the operators, and the limited number and difficulty of scenarios that were presented. These limitations should be evaluated with future work before the DST can move closer to real world implementation.

Future studies may benefit from evaluating a larger range of scenarios than what was evaluated in the aforementioned experiment. Specifically, the effect that the DST has on supervisor performance may not follow the results reported in Chapter 5 if the multiple alert scenarios had spacing between the alerts as opposed to both alerts occurring simultaneously.

There are many questions that arise from this work that could be studied further:

- Would DST users be able to correctly identify problematic scenarios if not provided an alert, or would automation bias result in detrimental performance for problematic scenarios that do not trigger an alert from the model (Type II errors)?
- Are some types of abnormal operator behaviors (such as low interaction frequency) easier for humans to identify than the HSMM-based operator model?
- Is the DST scalable to larger numbers of operators and more dynamic operations?
- How many operators would the ideal team consist of for a team supervisor using the DST?
- How far in the future are the HSMM-based operator model predictions beneficial?
- How would the same performance differ if the operators were allowed to see their own DST throughout the mission?
- Do subjects that are only given the alerts from the DST perform differently than subjects who are given the DST in its entirety (Interaction Frequency, Model Accuracy Prediction, and Confidence History plots)?
- How can the DST be adapted to account for a team of multiple, interdependent operators?

#### **6.4 Thesis Summary**

As the military and civilian sectors continue to invest in UVs, the scenario suggested in this thesis of a team supervisor coordinating the efforts of multiple, independent operators each controlling multiple heterogeneous UVs is likely (Cummings, et al., 2007; DoD, 2007b, 2009). Developing decision support tools in order aid the team supervisor in making time critical decisions that have possible catastrophic consequences is essential (Burns, 1978; Hackman, 2002; Leveson, 1986). The DST proposed in this thesis has demonstrated that using HSMM-based operator modeling to alert the team supervisor to current and predicted future operator abnormal behavior leads to increased team supervisor performance in terms of increased Decision Accuracy, decreased Incorrect Interventions, and decreased Response Times (single alert, true positive scenario only) in single alert scenarios. However, future work needs to be accomplished before the DST can move into real world production.





## References

- Bisantz, A. M., Kirlik, A., Gay, P., Phipps, D. A., Walker, N., & Fisk, A. D. (2000). Modeling and Analysis of a Dynamic Judgment Task Using a Lens Model Approach. *IEEE Transactions on Systems, Man, and Cybernetics*, 30(6), 605-616.
- Blanchard, B. S. & Fabrycky, W. J. (1998). *Systems Engineering and Analysis* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Boussemart, Y. & Cummings, M. L. (2008). *Behavioral Recognition and Prediction of an Operator Supervising Multiple Heterogeneous Unmanned Vehicles*. Paper presented at the Humans Operating Unmanned Systems, HUMOUS'08, Brest, France.
- Boussemart, Y., Fargeas, J. L., Cummings, M. L., & Roy, N. (2009). *Comparing Learning Techniques for Hidden Markov Models of Human Supervisory Control Behavior*. Paper presented at the AIAA Infotech@Aerospace '09 Conference, Seattle, Washington.
- Brewer, N. & Ridgway, T. (1998). Effects of Supervisory Monitoring on Productivity and Quality of Performance. *Journal of Experimental Psychology: Applied*, 4(3), 211-227.
- Broadbent, D. E. (1958). *Perception and Communication*. Oxford: Pergamon.
- Burns, J. M. (1978). *Leadership*. New York: Harper&Row.
- Cannon-Bowers, J. A., Salas, E., & Converse, S. A. (1993). Shared Mental Model in Expert Team Decision Making. In N. J. Castellan, *Individual and Group Decision Making: Current Issues*. Hillsdale, NJ: Erlbaum. 221-246.
- Castonia, R. W. (2009). *The Design of a HSMM-based Operator State Monitoring Display*. (HAL 2009-04). Cambridge, MA: MIT Humans and Automation Lab.
- Castonia, R. W. (2010). *The Design of a HSMM-based Operator State Modeling Display*. Paper presented at the AIAA Infotech@Aerospace 2010, Atlanta, GA.
- Crandall, B., Klein, G., & Hoffman, R. (2006). *Working Minds: A Practitioner's Guide to Cognitive Task Analysis*. Cambridge, MA: The MIT Press.
- Cucker, F. & Smale, S. (2002). On the Mathematical Foundations of Learning. *Bulletin of the American Mathematical Society*, 39(1), 1-49.
- Cummings, M. L. (2004). *Automation Bias in Intelligent Time Critical Decision Support Systems*. Paper presented at the AIAA 3rd Intelligent Systems Conference, Chicago, IL.
- Cummings, M. L., Bruni, S., Mercier, S., & Mitchell, P. J. (2007). Automation Architecture for Single Operator Multiple UAV Command and Control. *The International Command and Control Journal*, 1(2), 1-24.
- Cummings, M. L. & Guerlain, S. (2007). Developing Operator Capacity Estimates for Supervisory Control of Autonomous Vehicles. *Human Factors*, 49(1), 1-15.

- De Martino, B., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, Biases, and Rational Decision-Making in the Human Brain. *Science*, 313, 687-697.
- DoD. (2007a). *Joint Publication 1-02: DOD Dictionary of Military and Associated Terms*. Washington, D.C.: Office of the Joint Chiefs of Staff.
- DoD. (2007b). *Unmanned Systems Roadmap (2007-2032)*. Washington, D.C.: Office of the Secretary of Defense.
- DoD. (2009). *United States Air Force Unmanned Aircraft Systems Flight Plan*. Washington, D.C.: Office of the Secretary of the Air Force.
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The Role of Trust in Automation Reliance. *International Journal Human-Computer Studies*, 58(1), 697-718.
- Gardenier, J. S. (1981). Ship Navigational Failure Detection and Diagnosis. In J. Rasmussen & W. B. Rouse, *Human Detection and Diagnosis of System Failure*. Boston: Plenum. 49-74.
- Gegenfurtner, K. R. & Sharpe, L. T. (2001). *Color Vision: From Genes to Perception*: Cambridge University Press.
- Gigerenzer, G. & Goldstein, D. G. (1996). Reasoning the Fast and Frugal Way: Models of Bounded Rationality. *Psychological Review*, 103(4), 650-669.
- Guedon, Y. (2003). Estimating Hidden Semi-Markov Chains From Discrete Sequences. *Journal of Computational & Graphical Statistics*, 12(3), 604-639.
- Guerin, B. & Innes, J. (1993). *Social Facilitation*: Cambridge University Press, Editions de la Maison des Sciences de l'Homme.
- Guerlain, S. & Bullemer, P. (1996). *User-Initiated Notification: A Concept for Aiding the Monitoring Activities of Process Control Operators*. Paper presented at the Human Factors and Ergonomics Society 40th Annual Meeting, Philadelphia, PA.
- Gutwin, C. & Greenberg, S. (2004). The Importance of Awareness for Team Cognition in Distributed Collaboration. In E. Salas & S. M. Fiore, *Team Cognition: Understanding the Factors That Drive Process and Performance*. Washington, D.C.: American Psychological Association. 177-201.
- Hackman, J. R. (2002). *Leading Teams: Setting the Stage for Great Performances*. Boston, MA: Harvard Business School Press.
- Hancock, P. A. & Warm, J. S. (1989). A Dynamic Model of Stress in Sustained Attention. *Human Factors*, 31(5), 519-537.
- Huang, H. (2009). *Developing an Abstraction Layer for the Visualization of HSMM-Based Predictive Decision Support*. MIT M.Eng Thesis, Cambridge, MA.
- Irving, R. H., Higgins, C. A., & Safayeni, F. R. (1986). Computerized Performance Monitoring Systems: Use and Abuse. *Communications of the ACM*, 29(8), 794-801.

- Judd, T. S. & Kennedy, G. E. (2004). *More Sense from Audit Trails: Exploratory Sequential Data Analysis*. Paper presented at the Beyond the Comfort Zone: Proceedings of the 21st Annual Ascilite Conference, Perth, Australia.
- Kay, J., Maisonneuve, N., Yacef, K., & Zaiane, O. (2006). *Mining Patterns of Events in Students' Teamwork Data*. Paper presented at the Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems (ITS 2006), Jhongli, Taiwan.
- Kerstholt, J. H., Passenier, P. O., Houttuin, K., & Schuffel, H. (1996). The Effect of a Prior Probability and Complexity on Decision Making in a Supervisory Control Task. *Human Factors*, 38(1), 65-78.
- Klein, G. (1999). *Sources of Power: How People Make Decisions*. Cambridge, MA: The MIT Press.
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4), 541-551.
- Leveson, N. G. (1986). Software Safety: Why, What, and How. *Computing Surveys*, 18(2), 125-163.
- Levine, J. M. & Moreland, R. L. (1998). Small Groups. In D. T. Gilbert, S. T. Fiske & L. Gardner, *The Handbook of Social Psychology*. Boston: McGraw-Hill. 415-469.
- Maule, A. J. & Hockey, G. R. J. (1993). State, Stress, and Time Pressure *Time Pressure and Stress in Human Judgement and Decision Making*. New York: Plenum Press.
- Mitchell, P. J., Cummings, M. L., & Sheridan, T. B. (2004). *Human Supervisory Control Issues in Network Centric Warfare*. (HAL 2004-01). Cambridge, MA: MIT Humans and Automation Lab.
- Muir, B. M. & Moray, N. (1996). Trust in Automation. Part II. Experimental Studies of Trust and Human Intervention in a Process Control Simulation. *Ergonomics*, 39(3), 429-460.
- Nehme, C. E. (2009). *Modeling Human Supervisory Control in Heterogeneous Unmanned Vehicle Systems*. MIT PhD Thesis, Cambridge, MA.
- Nehme, C. E., Mekdeci, B., Crandall, J., & Cummings, M. L. (2008). The Impact of Heterogeneity on Operator Performance in Futuristic Unmanned Vehicle Systems. *The International C2 Journal*, 2(2).
- Nehme, C. E., Scott, S. D., Cummings, M. L., & Furusho, C. Y. (2006). *Generating Requirements for Futuristic Heterogeneous Unmanned Systems*. Paper presented at the Human Factors and Ergonomics Society, San Francisco, CA.
- Olson, G. M., Herbsleb, J. D., & Rueter, H. H. (1994). Characterizing the Sequential Structure of Interactive Behaviors through Statistical and Grammatical Techniques. *Human-Computer Interaction*, 9, 427-472.
- Parasuraman, R. & Riley, V. A. (1997). Humans and Automation: Use, Misuse, Disuse, and Abuse. *Human Factors*, 39(2), 230-253.
- Pentland, A. & Liu, A. (1995). *Towards Augmented Control Systems*. Paper presented at the IEEE Intelligent Vehicles '95, Detroit, MI.

- Pinelle, D., Gutwin, C., & Greenberg, S. (2003). Task Analysis for Groupware Usability Evaluation: Modeling Shared Workspace Tasks with the Mechanics of Collaboration. *ACM Transactions on Computer-Human Interaction*, 10(4), 281-311.
- Polvichai, J., Lewis, M., Scerri, P., & Sycara, K. (Eds.). (2006). *Using a Dynamic Neural Network to Model Team Performance for Coordination Algorithm Configuration and Reconfiguration of Large Multi-Agent Teams* (Vol. 16). New York: ASME Press Series.
- Rabiner, L. & Juang, B. (1986). An Introduction to Hidden Markov Models. *ASSP Magazine, IEEE*, 3(1), 4-16.
- Rasmussen, J. (1983). Skills, Rules, and Knowledge; Signals, Signs, and Symbols, and Other Distractions in Human Performance Models. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-13*(3), 257-266.
- Rhodes, B. J., Bomberger, N. A., Zandipour, M., Garagic, D., Stolzar, L. H., Dankert, J. R., Waxman, A. M., & Seibert, M. (2009). Automated Activity Pattern Learning and Monitoring Provide Decision Support to Supervisors of Busy Environments. *Intelligent Decision Technologies*, 3(1), 59-74.
- Rouse, W. B. (1983). *Systems Engineering Models of Human-Machine Interaction*. New York: North Holland.
- Salas, E., Dickinson, T. L., Converse, S. A., & Tannenbaum, S. I. (1992). Toward an Understanding of Team Performance and Training. In R. W. Swezey & E. Salas, *Teams: Their Training and Performance*. Norwood, NJ: Albex. 3-29.
- Sanderson, P. M. & Fisher, C. (1994). Exploratory Sequential Data Analysis. *Human-Computer Interaction*, 9(3), 247-250.
- Schmidt, D. K. (1978). A Queuing Analysis of the Air Traffic Controller's Workload. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(6), 492-498.
- Schraagen, J. M., Chipman, S., & Shalin, V. E. (2000). *Cognitive Task Analysis*. Mahwah, NJ: Erlbaum.
- Scott, S. D., Rico, A. E., Furusho, C. Y., & Cummings, M. L. (2007). *Aiding Team Supervision in Command and Control Operations with Large-Screen Displays*. Paper presented at the Human Systems Integration Symposium, Annapolis, MD.
- Scott, S. D., Sasangohar, F., & Cummings, M. L. (2009). *Investigating Supervisory-Level Activity Awareness Displays for Command and Control Operations*. Paper presented at the Human Systems Integration Symposium, Annapolis, MD.
- Scott, S. D., Wan, J., Sasangohar, F., & Cummings, M. L. (2008). *Mitigating Supervisory-level Interruptions in Mission Control Operations*. Paper presented at the 2nd International Conference on Applied Human Factors and Ergonomics, Las Vegas, NV.
- Sondik, E. J. (1971). *The Optimal Control of Partially Observable Markov Processes*. Stanford PhD Thesis, Palo Alto, CA.

- Terran, L. (1999). *Hidden Markov Models for Human Computer Interface Modeling*. Paper presented at the International Joint Conferences on Artificial Intelligence, Workshop on Learning About Users, Stockholm, Sweden.
- Tufte, E. R. (1983). *The Visual Display of Quantitative Information* (2nd ed.). Cheshire, CT: Graphics Press.
- Tversky, A. & Kahneman, D. (1974). Judgement Under Uncertainty: Heuristics and Biases. *Science*, 185, 1124-1131.
- Tversky, A. & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*(211), 453-458.
- Weinstein, N. & Klein, W. (1995). Resistance of Personal Risk Perceptions to Debiasing Interventions. *Health Psychology*, 14(2), 132-140.
- Welford, A. T. (1952). The Psychological Refractory Period and the Timing of High-Speed Performance - A Review and a Theory. *British Journal of Psychology*, 43, 2-19.
- Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The Cognitive Walkthrough Method: A Practitioner's Guide. In J. Nielsen & R. Mack, *Usability Inspection Methods*. New York: John Wiley & Sons.
- Wickens, C. D. & Hollands, J. G. (2000). *Engineering Psychology and Human Performance* (3rd ed.). Upper Saddle River, N.J.: Prentice Hall.
- Witten, I. H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufmann Publishers Inc.
- Yerkes, R. M. & Dodson, J. D. (1908). The Relation of Strength of Stimulus to Rapidity of Habit-Formation. *Journal of Comparative Neurology and Psychology*, 18, 459-482.



## Appendix A: Decision Ladders

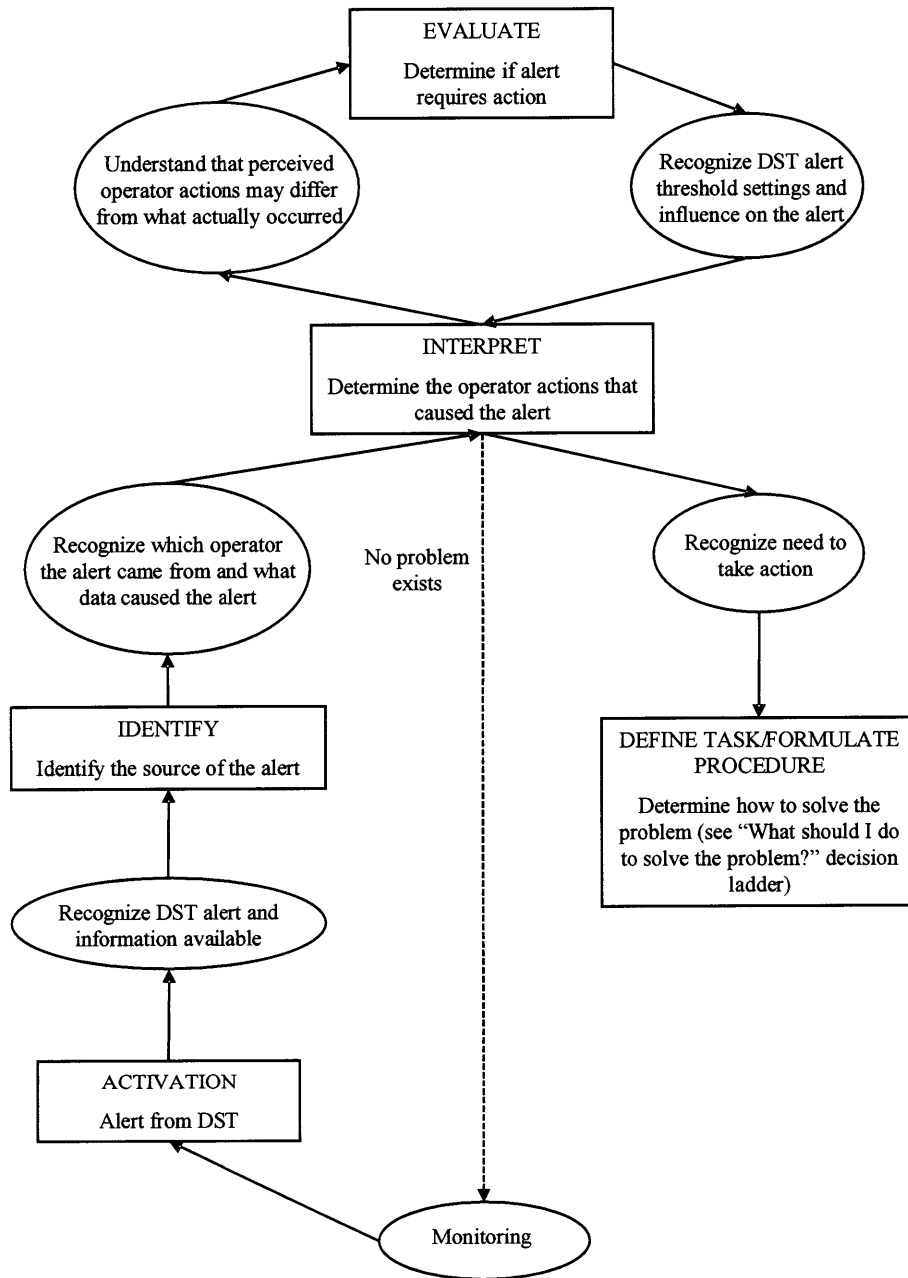


Figure A.1: "Is there a problem?" Decision Ladder

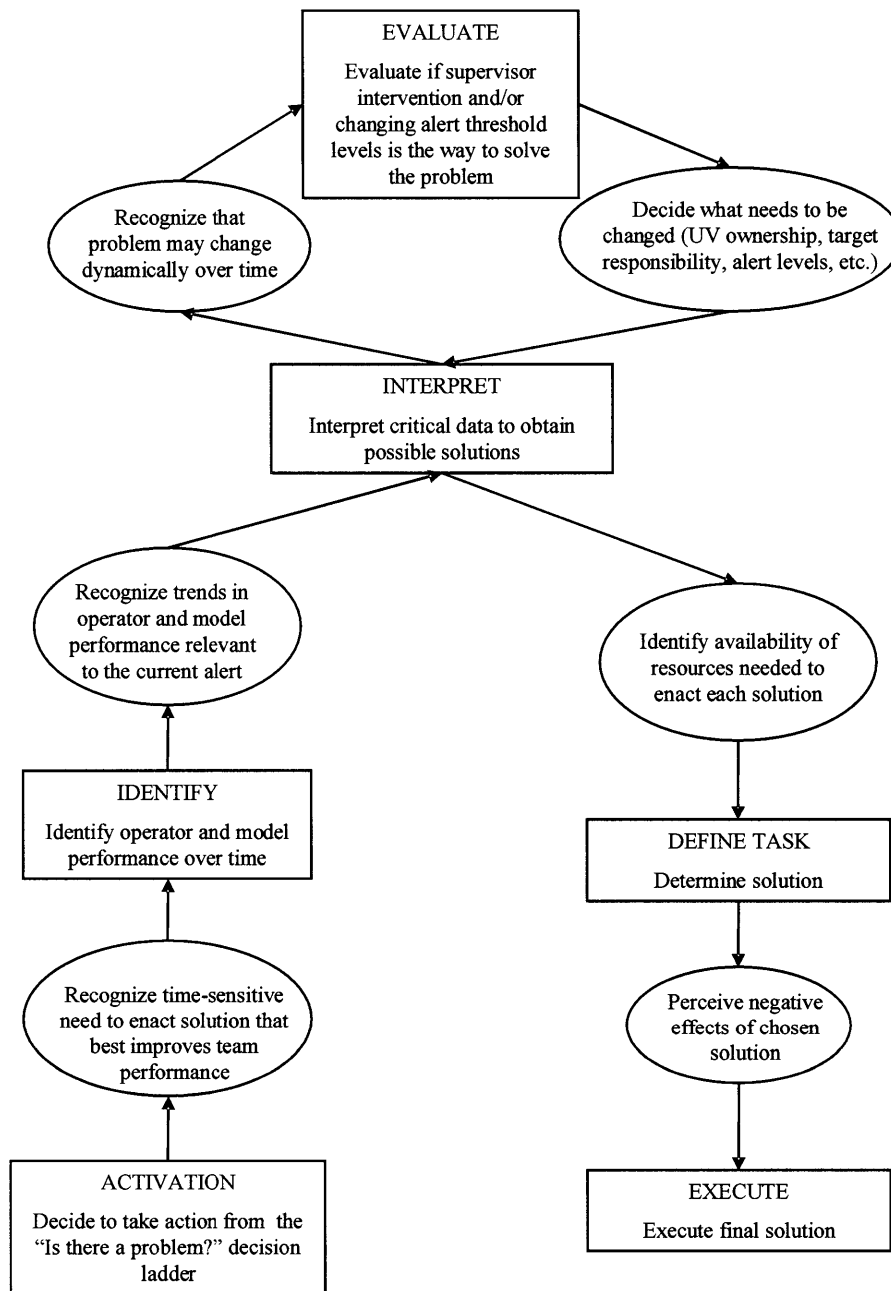


Figure A.2: "What should I do to solve the problem?" Decision Ladder



## Appendix B: Decision Ladders with Display Requirements for Decision Support Tool

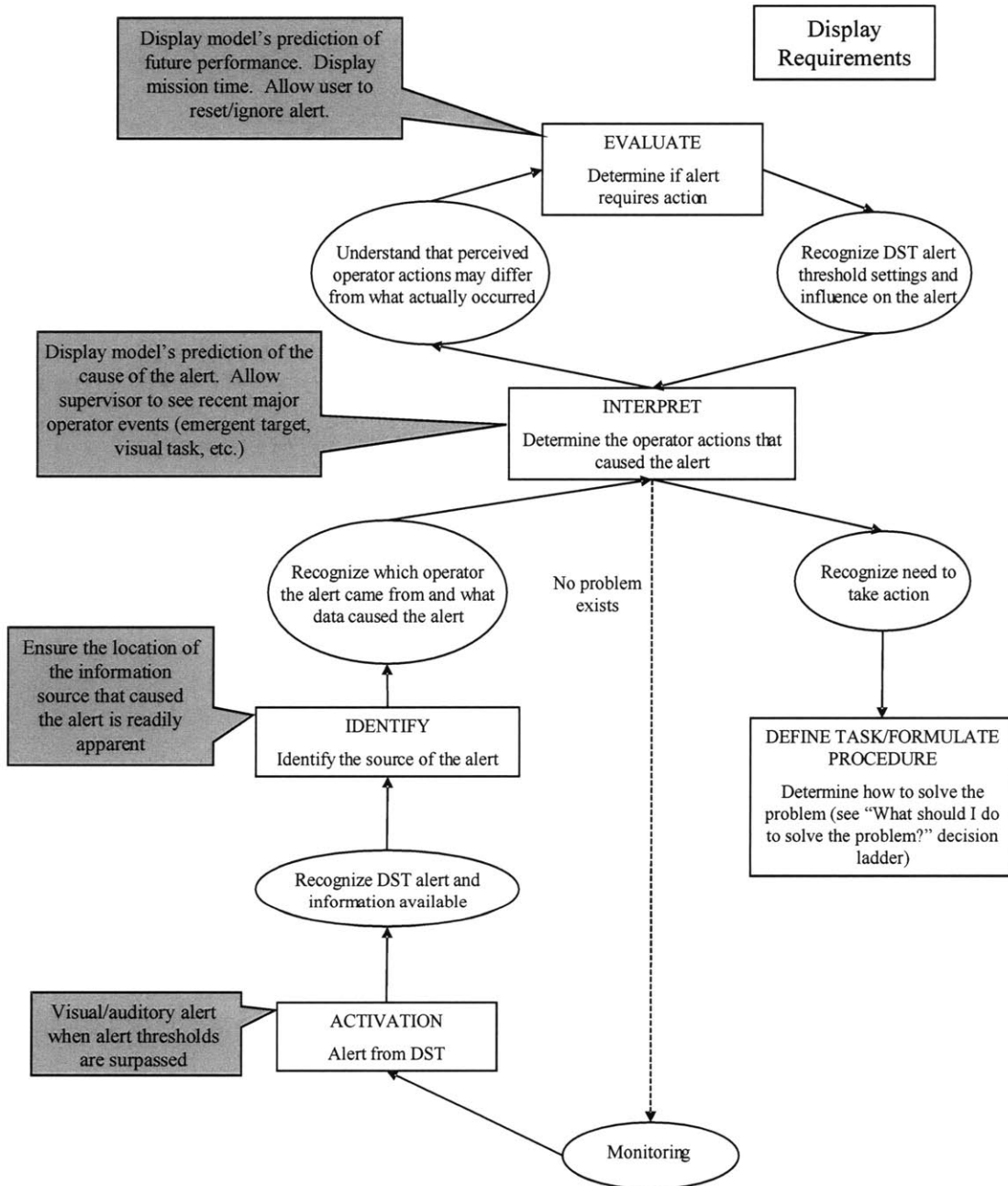


Figure B.1: "Is there a problem?" Decision Ladder with display requirements

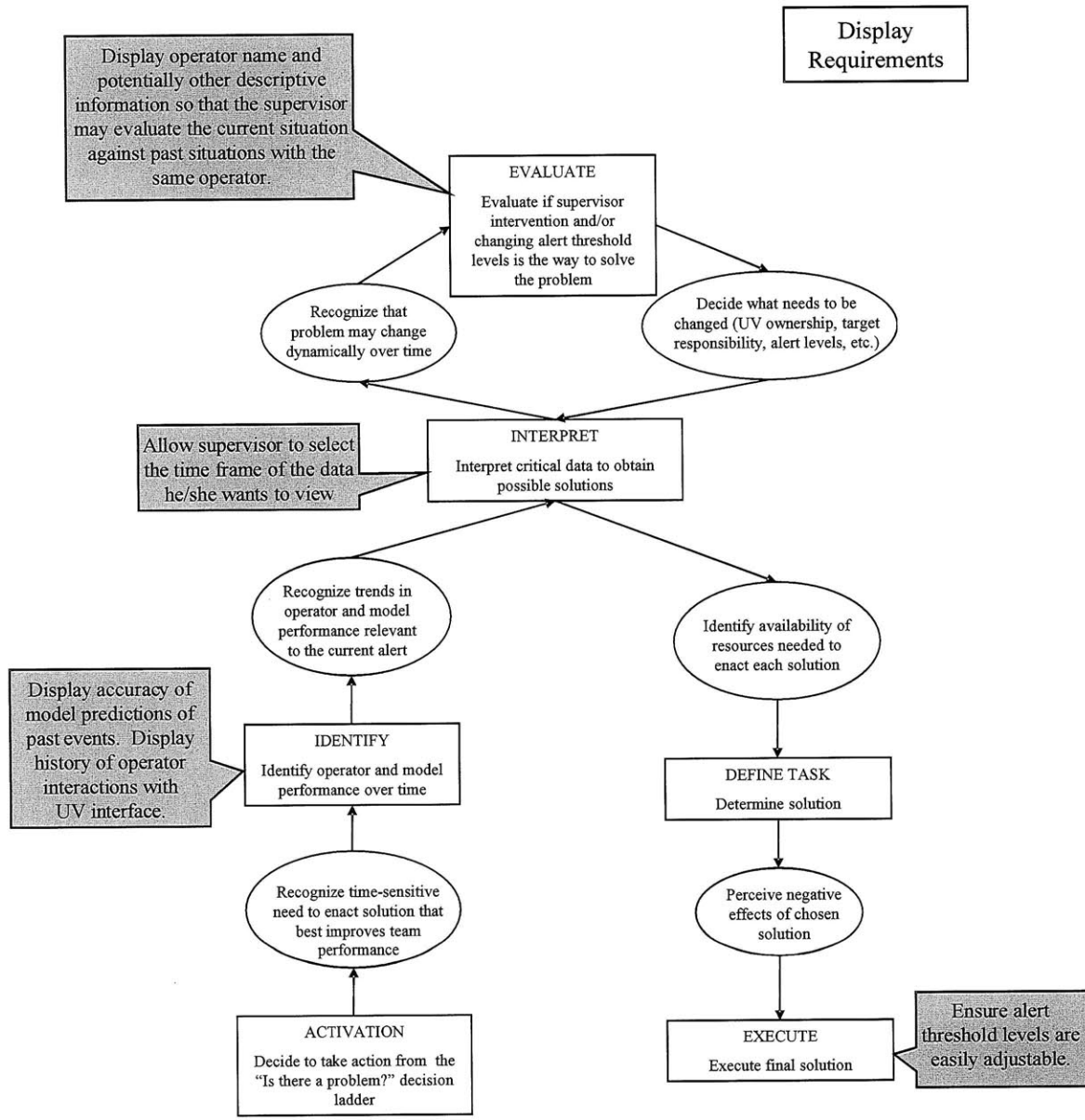


Figure B.2: "What should I do to solve the problem?" Decision Ladder with display requirements

## Appendix C: Scenario Descriptions and Test Matrix

	Scenario Description	Operator 1 Behavior	Operator 2 Behavior	Operator 3 Behavior
<b>Scenario 1</b>	true positive, single alert	normal	normal	True alert - Neglect HALEs (t=3:30)
<b>Scenario 2</b>	true positive, multiple alerts	True alert - low interaction frequency (t=3:20)	True alert - excessive status checking (t=3:20)	normal
<b>Scenario 3</b>	false positive, multiple alerts	normal	False alert - Neglect HALEs (t=2:15)	True alert - excessive waypoint modification (t=2:15)
<b>Scenario 4</b>	false positive, single alert	False alert - low interaction frequency (t=1:14)	normal	normal

Subject	Trial 1	Trial 2	Trial 3	Trial 4	DST
1	4	3	2	1	yes
2	1	2	4	3	yes
3	3	1	4	2	no
4	2	1	4	3	no
5	3	2	1	4	no
6	3	4	1	2	yes
7	4	2	3	1	no
8	3	4	2	1	yes
9	3	1	2	4	yes
10	1	3	2	4	yes
11	2	4	1	3	no
12	2	1	3	4	no
13	4	1	2	3	yes
14	3	2	4	1	no
15	4	1	3	2	no
16	1	4	2	3	no
17	1	3	4	2	no
18	2	3	1	4	yes
19	2	3	4	1	yes
20	1	2	3	4	yes
21	4	2	1	3	no
22	1	4	3	2	no
23	2	4	3	1	yes
24	4	3	1	2	yes
25	4	3	2	1	yes
26	1	2	4	3	yes
27	2	1	4	3	no
28	3	1	4	2	no
29	3	2	1	4	no
30	3	4	1	2	yes



## Appendix D: Human Subject Post Experiment Questionnaire

### Post Experiment Questionnaire

Use the following scale to answer questions 1-7 below.

- 1 – Strongly disagree
- 2 – Somewhat disagree
- 3 – Neutral
- 4 – Somewhat agree
- 5 – Strongly agree

	1	2	3	4	5
1. It was easy for me to understand the data presented in the Interaction Frequency plot.					
2. It was easy for me to understand the data presented in the Model Accuracy Prediction plot.					
3. It was easy for me to understand the data presented in the Confidence History plot.					
4. It was easy for me to understand the DST alerts.					
5. It was easy for me to understand the data presented in the DST as a whole.					
6. The information presented on the DST was beneficial in helping me make my decision of whether to intervene or not intervene.					
7. I primarily referenced the DST when making my decision to intervene or not intervene.					

Feel free to expand upon your answers and provide more details.

Questions 8-10 do not use the scale above.

8. What changes would you make to the DST to make it easier to use and why?

9. What information on the DST was most beneficial to you in deciding whether to intervene or not intervene?

10. Additional comments?



## **Appendix E: Human Subject Consent Form**

### **CONSENT TO PARTICIPATE IN NON-BIOMEDICAL RESEARCH**

Investigating Team Supervision Interfaces in Collaborative Time-Sensitive Targeting Operations

You are asked to participate in a research study conducted by Professor Mary Cummings Ph.D, (Principal Investigator) from the Aeronautics and Astronautics Department at the Massachusetts Institute of Technology (MIT) and Ryan Castonia (student investigator) from the Aeronautics and Astronautics Department at MIT. You were selected as a possible participant in this study because the expected population this research will impact is expected to contain men and women between the ages of 18 and 50 with an interest in using computers with possible military or military-in-training experience. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

#### **PARTICIPATION AND WITHDRAWAL**

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The student investigator may withdraw you from this research if circumstances arise which warrant doing so.

#### **PURPOSE OF THE STUDY**

This experiment will evaluate the effectiveness of a team supervisor real-time Decision Support Tool (DST) on improving team performance. This DST leverages recent advances in hidden semi-Markov model (HSMM) based operator state monitoring. The goals of this experiment are twofold. This first goal is to determine whether team supervisors with a HSMM-based operator modeling DST perform better (in terms of solving current problems efficiently and accurately, correctly preventing future problems from occurring, limiting unnecessary interventions, etc.) than supervisors without the DST. The second, more general goal is to address the question of how to best display probabilistic predictions and associated uncertainty, specifically the HSMM-based operator model output, so that someone with little to no background in statistics and probabilistic inference can efficiently and accurately make time-critical decisions.

#### **PROCEDURES**

If you volunteer to participate in this study, we would ask you to do the following things individually:

- Each participant begins by completing an informed consent form and a background questionnaire that gathers participants' demographic information. (~5 minutes)
- Attend training and practice session to learn a video game-like software environment that will have you monitoring the ongoing performance of a team of operators under your supervision and intervening when mission performance begins to degrade. Your team of operators (simulated in this experiment) will be

supervising and interacting with multiple unmanned vehicles to achieve the goals of your overall mission. (~20 minutes)

- Execute four trials consisting of the same tasks as above (~20 minutes).
- Fill out a post experiment questionnaire (~5 minutes).
- Attend a debrief session (~10 minutes).
- All testing will take place in MIT building 35, room 220.
- Total time: ~1 hr

#### **POTENTIAL RISKS AND DISCOMFORTS**

There are no anticipated physical or psychological risks in this study.

#### **POTENTIAL BENEFITS**

While there is no immediate foreseeable benefit to you as a participant in this study, your efforts will provide critical insight into the human cognitive capabilities and limitations for people who are expected to supervise multiple complex tasks at once, and how decision support visualizations can support their task management.

#### **PAYMENT FOR PARTICIPATION**

You will be paid \$10/hr for this effort today, and you will also have a chance to win a \$200 gift certificate. You will be notified at the completion of the study if you have won.

#### **CONFIDENTIALITY**

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law. You will be assigned a subject number which will be used on all related documents to include databases, summaries of results, etc. Only one master list of subject names and numbers will exist that will remain only in the custody of Professor Cummings.

#### **IDENTIFICATION OF INVESTIGATORS**

If you have any questions or concerns about the research, please feel free to contact the Principal Investigator, Mary L. Cummings, at (617) 252-1512, e-mail, [missyc@mit.edu](mailto:missyc@mit.edu), and her address is 77 Massachusetts Avenue, Room 33-311, Cambridge, MA, 02139. The student investigator is Ryan Castonia and may be contacted by telephone at (231) 740-1403 or via email at [Castonia@mit.edu](mailto:Castonia@mit.edu).

#### **EMERGENCY CARE AND COMPENSATION FOR INJURY**

In the unlikely event of physical injury resulting from participation in this research you may receive medical treatment from the M.I.T. Medical Department, including emergency treatment and follow-up care as needed. Your insurance carrier may be billed for the cost of such treatment. M.I.T. does not provide any other form of compensation for injury. Moreover, in either providing or making such medical care available it does not imply the injury is the fault of the investigator. Further information may be obtained by calling the MIT Insurance and Legal Affairs Office at 1-617-253-2822.



**RIGHTS OF RESEARCH SUBJECTS**

You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E32-335, 77 Massachusetts Ave, Cambridge, MA 02139, phone 1-617-253-6787.

**SIGNATURE OF RESEARCH SUBJECT OR LEGAL REPRESENTATIVE**

I understand the procedures described above and my questions have been answered to my satisfaction. I have been given a copy of this form.

I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to participate in this study. I have been given a copy of this form.

\_\_\_\_\_  
Name of Subject

\_\_\_\_\_  
Name of Legal Representative (if applicable)

\_\_\_\_\_  
Signature of Subject or Legal Representative

\_\_\_\_\_  
Date

**SIGNATURE OF INVESTIGATOR**

In my judgment the subject is voluntarily and knowingly giving informed consent and possesses the legal capacity to give informed consent to participate in this research study.

\_\_\_\_\_  
Signature of Investigator

\_\_\_\_\_  
Date



# Appendix F: Human Subject Demographic Survey

## *Collaborative TST Demographic Survey*

1. Age: \_\_\_\_\_

2. Gender:  Male  Female

3. Native Language: \_\_\_\_\_

**If native language is not English:**

*English Proficiency:*

Low

Moderate

High

4. Occupation: \_\_\_\_\_

**If student:**

a. *Class Standing:*  Undergraduate  Graduate

b. *Major:* \_\_\_\_\_

**If currently or formerly part of any country's armed forces:**

a. *Country/State:* \_\_\_\_\_

b. *Status:*  Active Duty  Reserve  Retired

c. *Service:*  Army  Navy  Air Force  Other \_\_\_\_\_

d. *Rank:* \_\_\_\_\_

e. *Years of Service:* \_\_\_\_\_

5. Have you had experience with remotely piloted vehicles (land, sea, air)?

Yes

No

**If yes:**

a. *Vehicle type(s)/class(es):*

\_\_\_\_\_

b. *Number of hours:* \_\_\_\_\_

**6. Have you had experience supervising a team of operators piloting vehicles (land, sea, air)?**

- Yes
- No

**If yes:**

b. *Vehicle type(s)/class(es):*

\_\_\_\_\_

c. *Responsibilities as team supervisor::* \_\_\_\_\_

d. *Size of teams:* \_\_\_\_\_

e. *Number of hours:* \_\_\_\_\_

**7. Do you have experience supervising a team of people in situations in which time was an important factor**

- Yes
- No

**If yes:**

f. *Types of situations in which time was an important factor:*

\_\_\_\_\_

g. *Responsibilities as team supervisor:*

\_\_\_\_\_

c. *Size of teams:* \_\_\_\_\_

d. *Number of hours:* \_\_\_\_\_

**8. Do you have experience supervising a team of people in other non time-critical situations**

- Yes
- No

**If yes:**

h. *Types of non time-critical situations:*

\_\_\_\_\_

i. *Responsibilities as team supervisor:*

\_\_\_\_\_

c. *Size of teams:* \_\_\_\_\_

d. *Number of hours:* \_\_\_\_\_

**9. How often do you play video games?**

- Never
- Less than 1 hour per week
- Between 1 and 4 hours per week
- Between 1 and 2 hours per day
- More than 2 hours per day

**10. Are you color blind?**


- Yes
- No

**If yes:**

Which type of color blindness (if known) \_\_\_\_\_



## Appendix G: Training Slides – DST User



Decision Support for Hidden Markov  
Model Predictions in Supervisory  
Control Settings

Ryan Castonia – S.M. Candidate  
Prof M.L. Cummings

MIT Humans and Automation Lab – Nov/Dec 09 1

### Overview

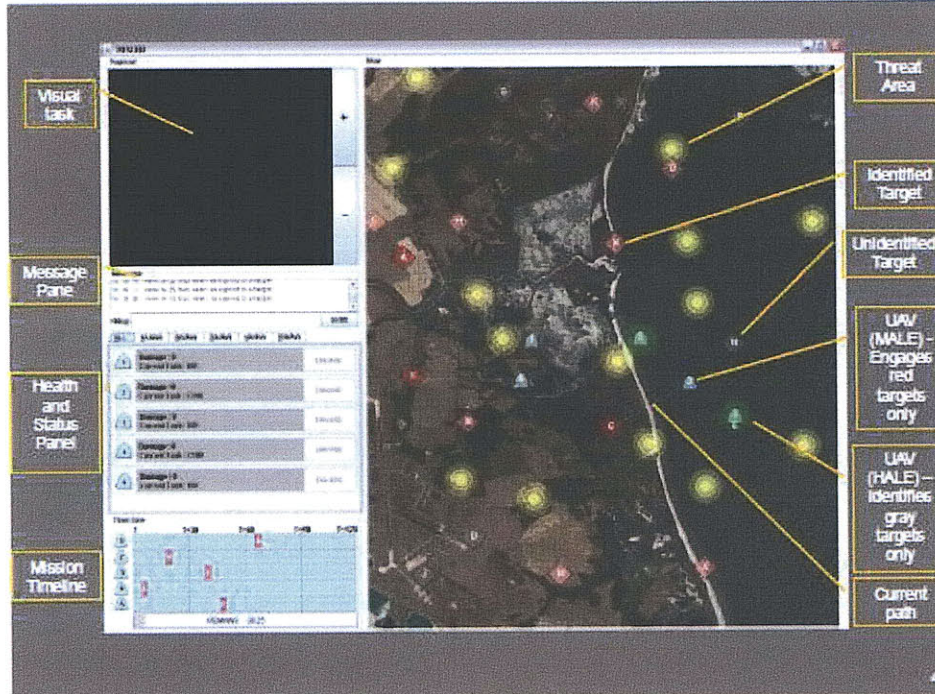
- You are a supervisor of a team of three operators who each control multiple unmanned aerial vehicles (UAVs)
- Your goal is to keep the team working as smoothly and efficiently as possible
- The instructional slides to follow will cover the following items:
  - Operator Interface (RESCHU)
  - Supervisor Interface
    - Decision Support Tool with predictive capability

2

# Operator Interface (RESCHU)

- Each operator controls multiple unmanned aerial vehicles (UAV)
- Each operator must perform several tasks: path planning, path re-planning due to emergent threat areas, reconnaissance of a visual task, and the evaluation of emergent targets
- You only need a basic understanding of what the operators are doing. You don't need to know the details behind how everything works.

3



4



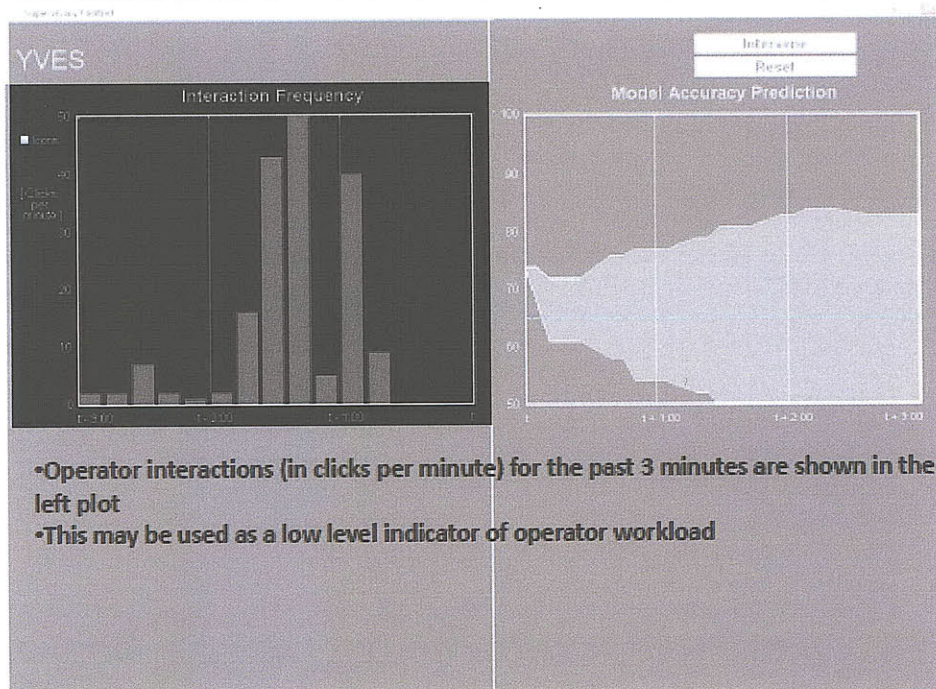
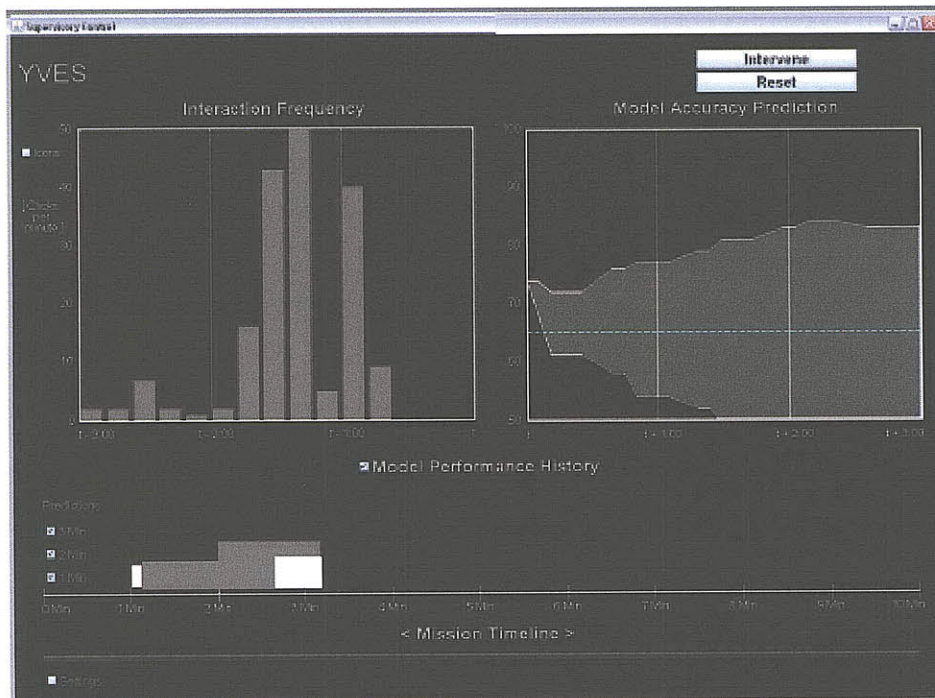
## Supervisor Interface

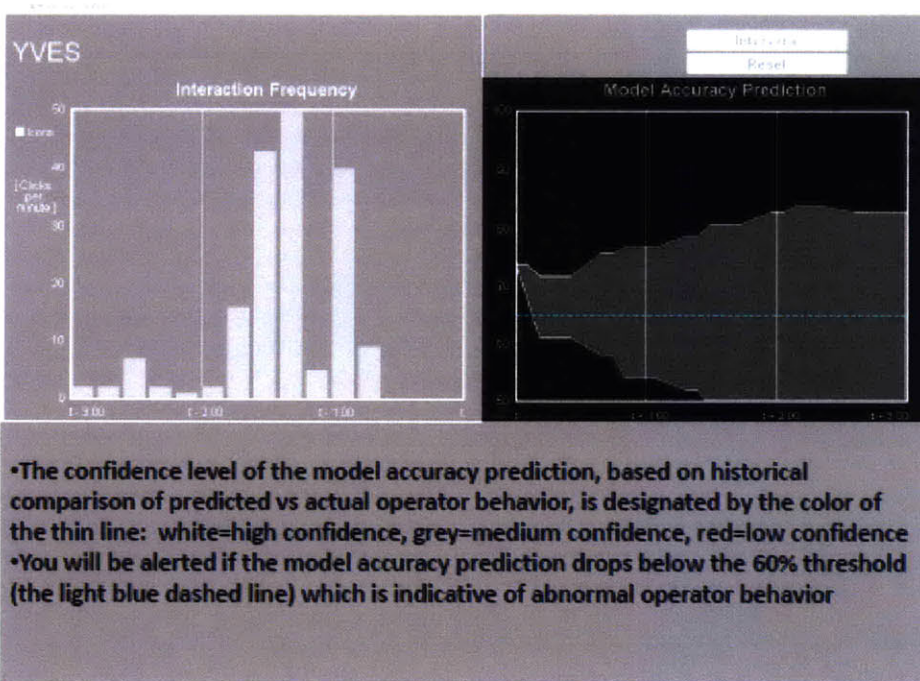
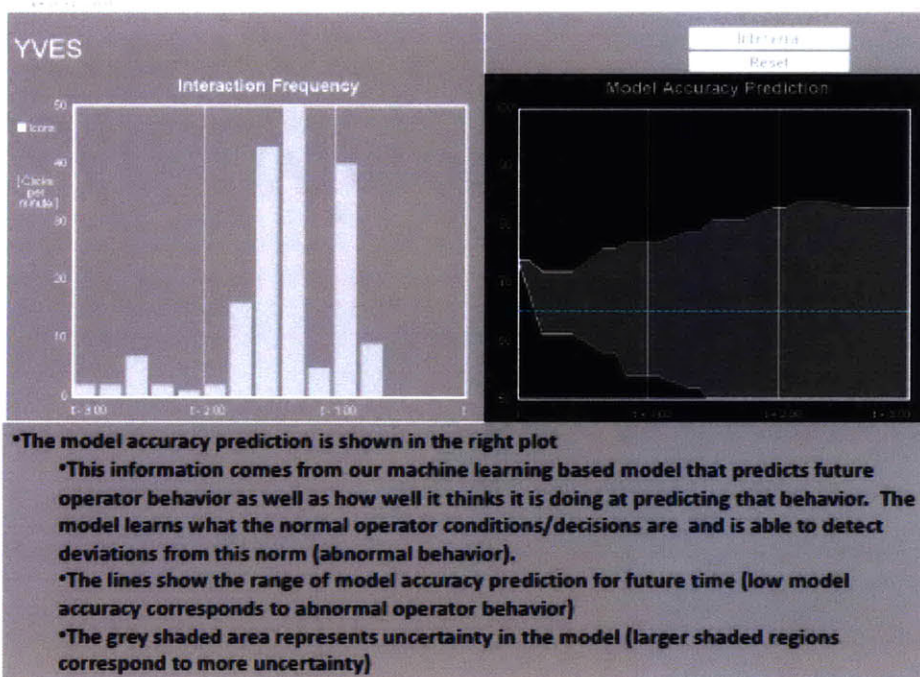
- Designed for supervisor of RESCHU operators
- Provides supervisor with prediction capabilities
  - Able to alert supervisor to present and predicted abnormal operator behavior so that the supervisor may prevent future problems from occurring
- Designed to help make two major decisions
  - Is there really a problem?
  - How do I solve the problem?

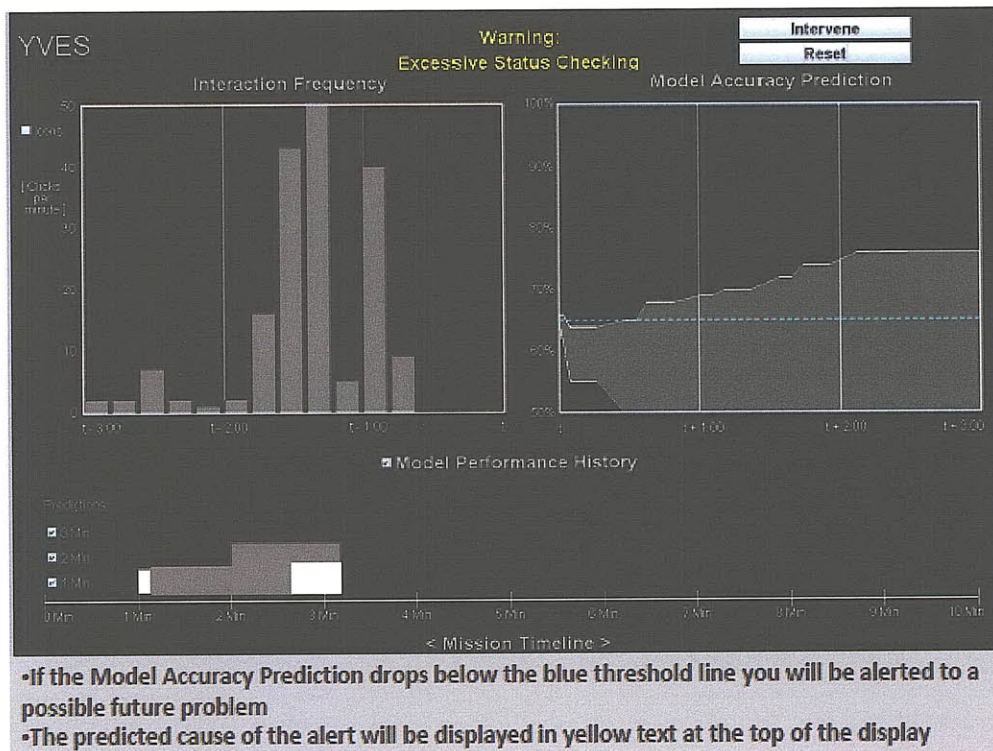
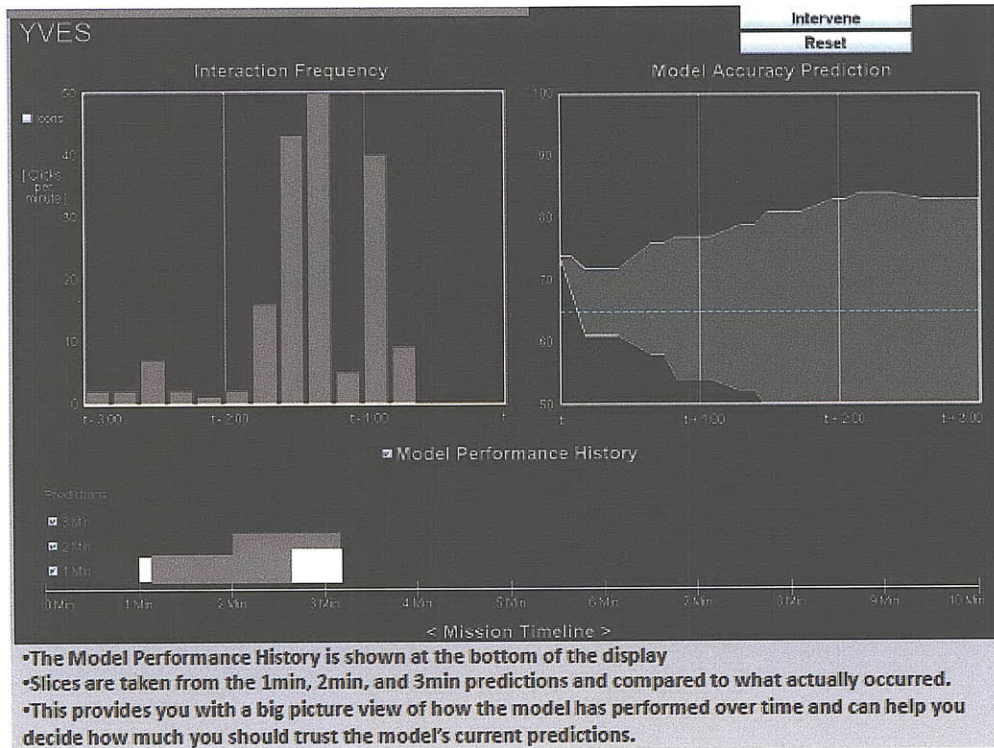
5

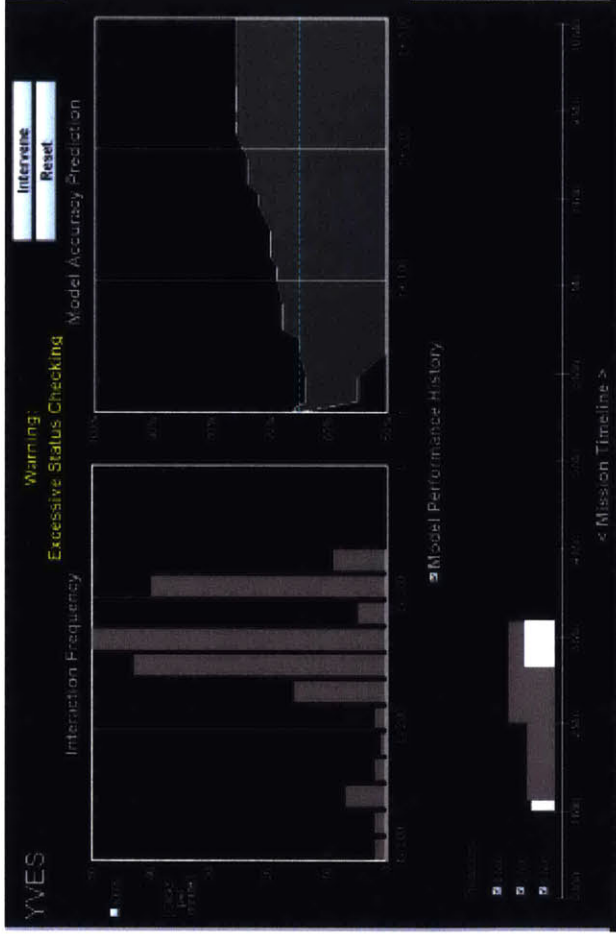
## Supervisor Interface

6

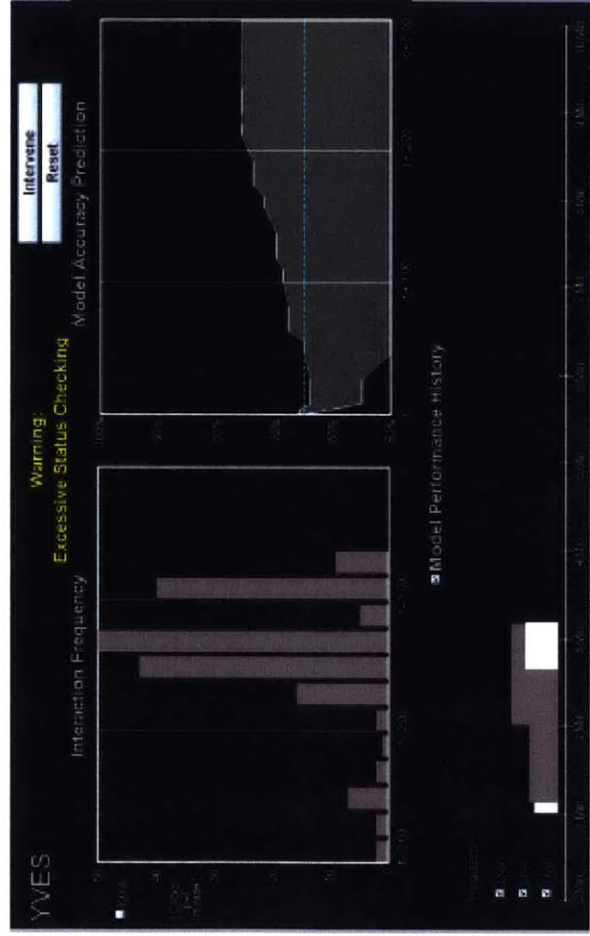








**In this example, the model accuracy prediction dropped below the threshold level of 60% and Yves is expected to be spending too much time checking the status of the UAVs.**



**-You must then use the information presented to decide whether to 'Intervene' in the situation or 'Reset' the alert**  
**-Intervening would attempt to correct inappropriate behavior by the operator, while resetting the alert would not interfere with the operator's behavior**

## Summary


- You are a supervisor of a team of three operators who each control multiple unmanned aerial vehicles (UAVs)
- Your goal is to keep the team working as smoothly and efficiently as possible (intervene when you deem necessary)
- The instructional slides covered the following items:
  - Operator Interface (RESCHU)
  - Supervisor Interface
    - Decision Support Tool with predictive capability

15

Questions?

16

## Appendix H: Training Slides – Non DST User



Decision Support for Hidden Markov  
Model Predictions in Supervisory  
Control Settings

Ryan Castonia – S.M. Candidate  
Prof M.L. Cummings

MIT Humans and Automation Lab – Nov/Dec 09 1

### Overview

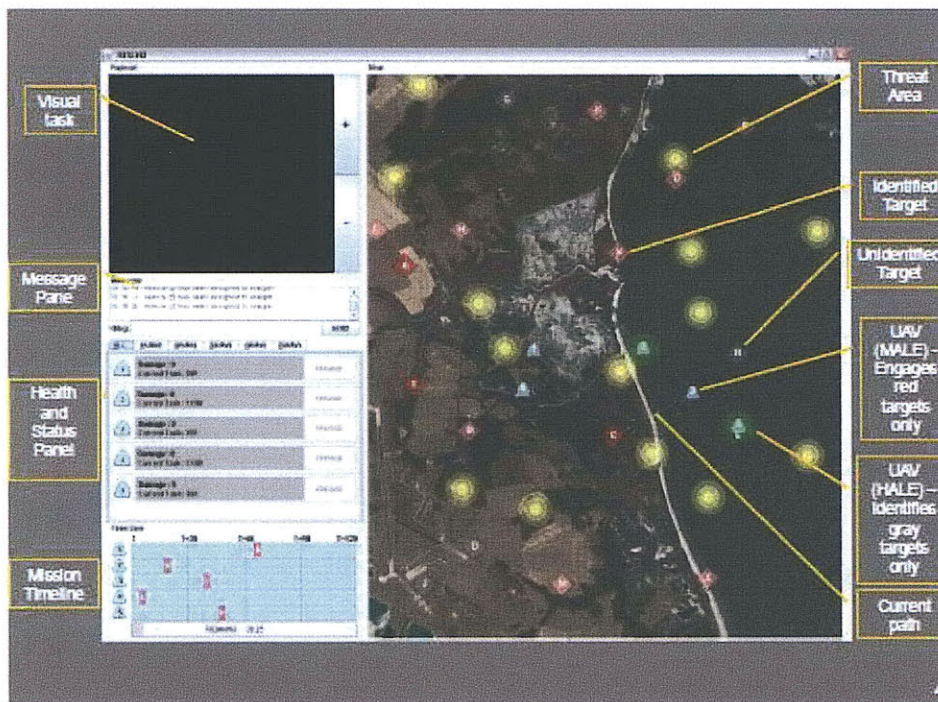
- You are a supervisor of a team of three operators who each control multiple unmanned aerial vehicles (UAVs)
- Your goal is to keep the team working as smoothly and efficiently as possible
- The instructional slides to follow will cover the following items:
  - Operator Interface (RESCHU)

2

## Operator Interface (RESCHU)

- Each operator controls multiple unmanned aerial vehicles (UAV)
- Each operator must perform several tasks: path planning, path re-planning due to emergent threat areas, reconnaissance of a visual task, and the evaluation of emergent targets
- You only need a basic understanding of what the operators are doing. You don't need to know the details behind how everything works.

3



4



## Operator Interface (RESCHU)

- Each operator controls multiple unmanned aerial vehicles (UAV)
- Each operator must perform several tasks: path planning, path re-planning due to emergent threat areas, reconnaissance of a visual task, and the evaluation of emergent targets
- You only need a basic understanding of what the operators are doing. You don't need to know the details behind how everything works.

5

## Summary

- You are a supervisor of a team of three operators who each control multiple unmanned aerial vehicles (UAVs)
- Your goal is to keep the team working as smoothly and efficiently as possible (intervene when you deem necessary)
- The instructional slides covered the following items:
  - Operator Interface (RESCHU)

6

Questions?

7

## Appendix I: Supporting Statistics

Numbers in parentheses correspond to Scenario numbers

**Table I.1: DST users, True Positive Single Alert Scenario descriptive statistics (1)**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	15	1	1	1.00	1	1	0.00
IncorrectInterventions	15	0	2	.40	.00	0	.74
ResponseTime1	15	1.3	12.6	6.22	6.2	-	2.88
SecondaryTaskRatio	15	.00	1.00	.37	.40	.40	.28

**Table I.2: Non-DST users, True Positive Single Alert Scenario descriptive statistics (1)**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	15	0	1	.47	0	0	.52
IncorrectInterventions	15	0	6	1.40	1	1	1.72
ResponseTime1	7	-10.0	26.0	13.14	20.0	-	12.59
SecondaryTaskRatio	15	.00	.80	.40	.40	.20	.24

**Table I.3: DST users, True Positive Multiple Alert Scenario descriptive statistics (2)**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	15	0	2	1.47	2	2	.64
IncorrectInterventions	15	0	2	.13	0	0	.52
ResponseTime1	15	-34.0	10.9	-3.81	1.2	-	14.15
ResponseTime2	12	-10.0	22.1	11.11	12.4	-	8.55
SecondaryTaskRatio	15	.00	.71	.34	.29	.29	.26
ResponseInterval	12	3.0	33.0	13.78	9.55	-	9.29

**Table I.4: Non-DST users, True Positive Multiple Alert Scenario descriptive statistics (2)**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	15	1	2	1.67	2	2	.49
IncorrectInterventions	15	0	2	.67	0	0	.90
ResponseTime1	15	-35.0	-4.0	-19.93	-20.0	-	9.01
ResponseTime2	10	-8.0	20.0	5.90	6.0	-	10.51
SecondaryTaskRatio	15	.14	.72	.37	.43	.43	.17
ResponseInterval	10	7.0	50.0	25.90	20.0	-	14.91

**Table I.5: DST users, False Positive Multiple Alert Scenario descriptive statistics (3)**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	15	0	2	1.53	2	2	.74
IncorrectInterventions	15	0	1	.07	0	0	.26
ResponseTime1	15	-8.0	16.9	6.52	6.8	-	5.50
ResponseTime2	15	10.0	32.0	19.71	17.5	-	6.75
SecondaryTaskRatio	15	.00	.50	.35	.50	.50	.18

**Table I.6: Non-DST users, False Positive Multiple Alert Scenario descriptive statistics (3)**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	15	1	2	1.60	2	2	.51
IncorrectInterventions	15	0	3	.73	1	0	.88
ResponseTime1	12	-15.0	18.0	3.72	5.00	-	10.13
ResponseTime2	1	7.0	7.0	7.0	7.0	-	-
SecondaryTaskRatio	15	.00	1.50	.35	.25	.25	.35

**Table I.7: DST users, False Positive Single Alert Scenario descriptive statistics (4)**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	15	1	1	1	1	1	0
IncorrectInterventions	15	0	0	0	0	0	0
ResponseTime1	15	2.3	14.6	6.00	5.5	-	3.21
SecondaryTaskRatio	15	.00	.83	.25	.17	.17	.22

**Table I.8: Non-DST users, False Positive Single Alert Scenario descriptive statistics (4)**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	15	0	1	.93	1	1	.26
IncorrectInterventions	15	0	1	.20	0	0	.41
ResponseTime1	1	3.0	3.0	3.0	3.0	-	-
SecondaryTaskRatio	15	.00	.83	.21	.17	.17	.22

**Table I.9: DST users, True Positive Scenarios descriptive statistics (1 and 2)**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	30	0	2	1.23	1	1	.50
IncorrectInterventions	30	0	2	.27	0	0	.64
ResponseTime1	30	-34.0	12.6	1.20	6.0	6.6	11.25
SecondaryTaskRatio	30	.00	1.00	.36	.35	.00	.27

**Table I.10: Non-DST users, True Positive Scenarios descriptive statistics (1 and 2)**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	30	0	2	1.07	1	1	.79
IncorrectInterventions	30	0	6	1.03	1	0	1.40
ResponseTime1	22	-35.0	26.0	-9.41	-15.0	-	18.66
SecondaryTaskRatio	30	.00	.80	.39	.42	.20	.20

**Table I.11: DST users, False Positive Scenarios descriptive statistics (3 and 4)**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	30	0	2	1.27	1	1	.58
IncorrectInterventions	30	0	1	.03	0	0	.18
ResponseTime1	30	-8.0	16.9	6.26	6.0	-	4.43
SecondaryTaskRatio	30	.00	.83	.30	.25	.50	.21

**Table I.12: Non-DST users, False Positive Scenarios descriptive statistics (3 and 4)**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	30	0	2	1.27	1	1	.52
IncorrectInterventions	30	0	3	.47	0	0	.73
ResponseTime1	13	-15.0	18.0	3.85	5.0	-	9.70
SecondaryTaskRatio	30	.00	1.50	.28	.25	.25	.30

**Table I.13: DST users, Single Alert Scenarios descriptive statistics (1 and 4)**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	30	1	1	1.00	1	1	.00
IncorrectInterventions	30	0	2	.20	0	0	.55
ResponseTime1	30	1.3	14.6	6.1	5.8	-	3.00
SecondaryTaskRatio	30	.00	1.00	.31	.20	.17	.26

**Table I.14: Non-DST users, Single Alert Scenarios descriptive statistics (1 and 4)**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	30	0	1	.70	1	1	.47
IncorrectInterventions	30	0	6	.80	0	0	1.38
ResponseTime1	8	-10.0	26.0	11.9	14.0	-	12.20
SecondaryTaskRatio	30	.00	.83	.31	.20	.17	.25

**Table I.15: DST users, Multiple Alert Scenarios descriptive statistics (2 and 3)**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	30	0	2	1.50	2	2	.68
IncorrectInterventions	30	0	2	.10	0	0	.40
ResponseTime1	30	-34.0	16.9	1.35	6.55	-	11.78
ResponseTime2	27	-10.0	32.0	15.89	16.3	-	8.63
SecondaryTaskRatio	30	.00	.71	.35	.29	.50	.22
ResponseInterval	12	3.0	33.0	13.8	9.6	-	9.29

**Table I.16: Non-DST users, Multiple Alert Scenarios descriptive statistics (2 and 3)**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	30	1	2	1.63	2	2	.49
IncorrectInterventions	30	0	3	.70	0	0	.88
ResponseTime1	27	-35.0	18.0	-9.3	-11.0	-	15.26
ResponseTime2	11	-8.0	20.0	6.0	7.0	-	9.98
SecondaryTaskRatio	30	.00	1.50	.36	.29	.25	.27
ResponseInterval	10	7.0	50.0	25.9	20.0	-	14.91

**Table I.17: DST Users, All Scenarios descriptive statistics**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	60	0	2	1.25	1	1	.54
IncorrectInterventions	60	0	2	.15	0	0	.48
ResponseTime1	60	-34.0	16.9	3.7	6.1	-	8.85
ResponseTime2	27	-10.0	32.0	15.9	16.3	-	8.63
SecondaryTaskRatio	60	.00	1.00	.33	.29	.00	.24
ResponseInterval	12	3.0	33.0	13.8	9.6	-	9.29

**Table I.18: Non-DST Users, All Scenarios descriptive statistics**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	60	0	2	1.17	1	1	.67
IncorrectInterventions	60	0	6	.75	0	0	1.14
ResponseTime1	35	-35.0	26.0	-4.5	-5.0	-	17.04
ResponseTime2	11	-8.0	20.0	6.0	7.0	-	9.98
SecondaryTaskRatio	60	.00	1.50	.33	.25	.25	.26
ResponseInterval	10	7.0	50.0	25.9	20.0	-	14.91

**Table I.19: All Subjects, All Scenarios descriptive statistics**

	N	Minimum	Maximum	Mean	Median	Mode	Std. Deviation
CorrectDecision	120	0	2	1.21	1	1	.61
IncorrectInterventions	120	0	6	.45	0	0	.92
ResponseTime1	95	-35.0	26.0	.7	5.2	-	13.04
ResponseTime2	38	-10.0	32.0	13.0	14.7	-	10.00
SecondaryTaskRatio	120	.00	1.50	.33	.27	.00	.25
ResponseInterval	22	3.0	50.0	19.3	15.0	-	13.37

**Table I.20: Normality Tests for DST vs non-DST distributions**

**Tests of Normality**

	DST	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
CorrectDecision	no	.282	60	.000	.791	60	.000
	yes	.378	60	.000	.709	60	.000
IncorrectInterventions	no	.294	60	.000	.680	60	.000
	yes	.522	60	.000	.344	60	.000
ResponseTime1	no	.103	35	.200*	.963	35	.287
	yes	.250	60	.000	.729	60	.000
ResponseTime2	no	.138	11	.200*	.933	11	.439
	yes	.122	27	.200*	.953	27	.248
SecondaryTaskRatio	no	.167	60	.000	.863	60	.000
	yes	.131	60	.012	.942	60	.007
ResponseInterval	no	.254	10	.067	.900	10	.219
	yes	.197	11	.200*	.896	11	.165

\*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

**Table I.21: Homogeneity of Variance Tests for DST vs non-DST distributions**

**Robust Tests of Equality of Means**

		Statistic <sup>a</sup>	df1	df2	Sig.
CorrectDecision	Brown-Forsythe	.564	1	113.088	.454
IncorrectInterventions	Brown-Forsythe	14.020	1	79.219	.000
ResponseTime1	Brown-Forsythe	7.029	1	44.908	.011
ResponseTime2	Brown-Forsythe	8.274	1	16.427	.011
SecondaryTaskRatio	Brown-Forsythe	.017	1	117.275	.898
ResponseInterval	Brown-Forsythe	4.995	1	14.526	.042

a. Asymptotically F distributed.

**Table I.22: Kruskal-Wallis tests for ordering effects**

**Order=1, first Scenario presented to subject**

**Order=2, second Scenario presented to subject**

**Order=3, third Scenario presented to subject**

**Order=4, fourth Scenario presented to subject**

**Ranks**

	Order	N	Mean Rank
CorrectDecision	1	30	66.42
	2	30	58.67
	3	30	60.05
	4	30	56.87
	Total	120	
IncorrectInterventions	1	30	57.45
	2	30	59.90
	3	30	64.22
	4	30	60.43
	Total	120	
ResponseTime1	1	24	48.52
	2	23	42.76
	3	23	56.46
	4	25	44.54
	Total	95	
ResponseTime2	1	9	21.72
	2	12	20.46
	3	8	23.81
	4	9	12.17
	Total	38	
SecondaryTaskRatio	1	30	56.78
	2	30	56.90
	3	30	64.18
	4	30	64.13
	Total	120	

**Test Statistics<sup>a,b</sup>**

	Correct Decision	Incorrect Interventions	Response Time1	Response Time2	Secondary TaskRatio
Chi-Square	1.683	.948	3.399	5.578	1.341
df	3	3	3	3	3
Asymp. Sig.	.641	.814	.334	.134	.720

a. Kruskal Wallis Test

b. Grouping Variable: Order



DST=0, non-DST user

DST=1, DST user

**Table I.23: Mann-Whitney U Tests –True Positive Single Alert Scenario (1)**

**Ranks**

	DST	N	Mean Rank	Sum of Ranks
CorrectDecision	0	15	11.50	172.50
	1	15	19.50	292.50
	Total	30		
IncorrectInterventions	0	15	18.57	278.50
	1	15	12.43	186.50
	Total	30		
ResponseTime1	0	7	15.21	106.50
	1	15	9.77	146.50
	Total	22		
Secondary TaskRatio	0	15	16.17	242.50
	1	15	14.83	222.50
	Total	30		

**Test Statistics<sup>b</sup>**

	Correct Decision	Incorrect Interventions	Response Time1	Secondary TaskRatio
Mann-Whitney U	52.500	66.500	26.500	102.500
Wilcoxon W	172.500	186.500	146.500	222.500
Z	-3.247	-2.095	-1.834	-.430
Asymp. Sig. (2-tailed)	.001	.036	.067	.667
Exact Sig. [2*(1-tailed Sig.)]	.011 <sup>a</sup>	.056 <sup>a</sup>	.066 <sup>a</sup>	.683 <sup>a</sup>

a. Not corrected for ties.

b. Grouping Variable: DST\_

DST=0, non-DST user

DST=1, DST user

Table I.24: Mann-Whitney U Tests – True Positive Multiple Alert Scenario (2)

Ranks				
	DST	N	Mean Rank	Sum of Ranks
CorrectDecision	0	15	16.67	250.00
	1	15	14.33	215.00
	Total	30		
IncorrectInterventions	0	15	17.93	269.00
	1	15	13.07	196.00
	Total	30		
ResponseTime1	0	15	10.67	160.00
	1	15	20.33	305.00
	Total	30		
ResponseTime2	0	10	10.00	100.00
	1	12	12.75	153.00
	Total	22		
Secondary TaskRatio	0	15	16.13	242.00
	1	15	14.87	223.00
	Total	30		
ResponseInterval	0	10	14.65	146.50
	1	12	8.88	106.50
	Total	22		

Test Statistics <sup>b</sup>						
	Correct Decision	Incorrect Interventions	Response Time1	Response Time2	Secondary TaskRatio	Response Interval
Mann-Whitney U	95.000	76.000	40.000	45.000	103.000	28.500
Wilcoxon W	215.000	196.000	160.000	100.000	223.000	106.500
Z	-.846	-2.051	-3.018	-.990	-.401	-2.079
Asymp. Sig. (2-tailed)	.397	.040	.003	.322	.688	.038
Exact Sig. [2*(1-tailed Sig.)]	.486 <sup>a</sup>	.137 <sup>a</sup>	.002 <sup>a</sup>	.346 <sup>a</sup>	.713 <sup>a</sup>	.036 <sup>a</sup>

a. Not corrected for ties.

b. Grouping Variable: DST\_

DST=0, non-DST user

DST=1, DST user

**Table I.25: Mann-Whitney U Tests – False Positive Multiple Alert Scenario (3)**

**Ranks**

	DST	N	Mean Rank	Sum of Ranks
CorrectDecision	0	15	15.40	231.00
	1	15	15.60	234.00
	Total	30		
IncorrectInterventions	0	15	19.07	286.00
	1	15	11.93	179.00
	Total	30		
ResponseTime1	0	12	13.17	158.00
	1	15	14.67	220.00
	Total	27		
Secondary TaskRatio	0	15	14.03	210.50
	1	15	16.97	254.50
	Total	30		

**Test Statistics<sup>b</sup>**

	Correct Decision	Incorrect Interventions	Response Time1	Secondary TaskRatio
Mann-Whitney U	111.000	59.000	80.000	90.500
Wilcoxon W	231.000	179.000	158.000	210.500
Z	-.073	-2.763	-.489	-.991
Asymp. Sig. (2-tailed)	.942	.006	.625	.322
Exact Sig. [2*(1-tailed Sig.)]	.967 <sup>a</sup>	.026 <sup>a</sup>	.648 <sup>a</sup>	.367 <sup>a</sup>

a. Not corrected for ties.

b. Grouping Variable: DST\_

DST=0, non-DST user

DST=1, DST user

**Table I.26: Mann-Whitney U Tests – False Positive Single Alert Scenario (4)**

**Ranks**

	DST	N	Mean Rank	Sum of Ranks
CorrectDecision	0	15	15.00	225.00
	1	15	16.00	240.00
	Total	30		
IncorrectInterventions	0	15	17.00	255.00
	1	15	14.00	210.00
	Total	30		
ResponseTime1	0	1	2.00	2.00
	1	15	8.93	134.00
	Total	16		
Secondary TaskRatio	0	15	14.73	221.00
	1	15	16.27	244.00
	Total	30		

**Test Statistics<sup>b</sup>**

	Correct Decision	Incorrect Interventions	Response Time1	Secondary TaskRatio
Mann-Whitney U	105.000	90.000	1.000	101.000
Wilcoxon W	225.000	210.000	2.000	221.000
Z	-1.000	-1.795	-1.411	-.508
Asymp. Sig. (2-tailed)	.317	.073	.158	.612
Exact Sig. [2*(1-tailed Sig.)]	.775 <sup>a</sup>	.367 <sup>a</sup>	.250 <sup>a</sup>	.653 <sup>a</sup>

a. Not corrected for ties.

b. Grouping Variable: DST\_

DST=0, non-DST user

DST=1, DST user

**Table I.27: Mann-Whitney U Tests – Single Alert Scenarios (1 and 4)**

**Ranks**

	DST	N	Mean Rank	Sum of Ranks
CorrectDecision	0	30	26.00	780.00
	1	30	35.00	1050.00
	Total	60		
IncorrectInterventions	0	30	34.93	1048.00
	1	30	26.07	782.00
	Total	60		
ResponseTime1	0	8	25.06	200.50
	1	30	18.02	540.50
	Total	38		
Secondary TaskRatio	0	30	30.77	923.00
	1	30	30.23	907.00
	Total	60		

**Test Statistics<sup>b</sup>**

	Correct Decision	Incorrect Interventions	Response Time1	Secondary TaskRatio
Mann-Whitney U	315.000	317.000	75.500	442.000
Wilcoxon W	780.000	782.000	540.500	907.000
Z	-3.227	-2.486	-1.594	-.120
Asymp. Sig. (2-tailed)	.001	.013	.111	.905
Exact Sig. [2*(1-tailed Sig.)]			.112 <sup>a</sup>	

a. Not corrected for ties.

b. Grouping Variable: DST\_

**DST=0, non-DST user**

**DST=1, DST user**

**Table I.28: Mann-Whitney U Tests – Multiple Alert Scenarios (2 and 3)**

**Ranks**

	DST	N	Mean Rank	Sum of Ranks
CorrectDecision	0	30	31.55	946.50
	1	30	29.45	883.50
	Total	60		
IncorrectInterventions	0	30	36.48	1094.50
	1	30	24.52	735.50
	Total	60		
ResponseTime1	0	27	22.94	619.50
	1	30	34.45	1033.50
	Total	57		
Secondary TaskRatio	0	30	29.38	881.50
	1	30	31.62	948.50
	Total	60		

**Test Statistics<sup>a</sup>**

	Correct Decision	Incorrect Interventions	Response Time1	Secondary TaskRatio
Mann-Whitney U	418.500	270.500	241.500	416.500
Wilcoxon W	883.500	735.500	619.500	881.500
Z	-.546	-3.422	-2.616	-.502
Asymp. Sig. (2-tailed)	.585	.001	.009	.616

a. Grouping Variable: DST\_

**DST=0, non-DST user**

**DST=1, DST user**

**Table I.29: Mann-Whitney U Tests – True Positive Scenarios (1 and 2)**

**Ranks**

	DST	N	Mean Rank	Sum of Ranks
CorrectDecision	0	30	28.90	867.00
	1	30	32.10	963.00
	Total	60		
IncorrectInterventions	0	30	36.02	1080.50
	1	30	24.98	749.50
	Total	60		
ResponseTime1	0	22	20.73	456.00
	1	30	30.73	922.00
	Total	52		
Secondary TaskRatio	0	30	32.00	960.00
	1	30	29.00	870.00
	Total	60		

**Test Statistics<sup>a</sup>**

	Correct Decision	Incorrect Interventions	Response Time1	Secondary TaskRatio
Mann-Whitney U	402.000	284.500	203.000	405.000
Wilcoxon W	867.000	749.500	456.000	870.000
Z	-.792	-2.886	-2.354	-.670
Asymp. Sig. (2-tailed)	.429	.004	.019	.503

a. Grouping Variable: DST\_

**DST=0, non-DST user**

**DST=1, DST user**

**Table I.30: Mann-Whitney U Tests – False Positive Scenarios (3 and 4)**

**Ranks**

	DST	N	Mean Rank	Sum of Ranks
CorrectDecision	0	30	30.37	911.00
	1	30	30.63	919.00
	Total	60		
IncorrectInterventions	0	30	35.53	1066.00
	1	30	25.47	764.00
	Total	60		
ResponseTime1	0	13	19.96	259.50
	1	30	22.88	686.50
	Total	43		
Secondary TaskRatio	0	30	28.72	861.50
	1	30	32.28	968.50
	Total	60		

**Test Statistics<sup>b</sup>**

	Correct Decision	Incorrect Interventions	Response Time1	Secondary TaskRatio
Mann-Whitney U	446.000	299.000	168.500	396.500
Wilcoxon W	911.000	764.000	259.500	861.500
Z	-.070	-3.211	-.701	-.809
Asymp. Sig. (2-tailed)	.944	.001	.483	.418
Exact Sig. [2*(1-tailed Sig.)]			.488 <sup>a</sup>	

a. Not corrected for ties.

b. Grouping Variable: DST\_



**DST=0, non-DST user**

**DST=1, DST user**

**Table I.31: Mann-Whitney U Tests – All Scenarios**

**Ranks**

	DST	N	Mean Rank	Sum of Ranks
CorrectDecision	0	60	58.88	3532.50
	1	60	62.13	3727.50
	Total	120		
IncorrectInterventions	0	60	70.93	4255.50
	1	60	50.08	3004.50
	Total	120		
ResponseTime1	0	35	39.20	1372.00
	1	60	53.13	3188.00
	Total	95		
ResponseTime2	0	11	12.50	137.50
	1	27	22.35	603.50
	Total	38		
Secondary TaskRatio	0	60	60.35	3621.00
	1	60	60.65	3639.00
	Total	120		

**Test Statistics<sup>b</sup>**

	Correct Decision	Incorrect Interventions	Response Time1	Response Time2	Secondary TaskRatio
Mann-Whitney U	1702.500	1174.500	742.000	71.500	1791.000
Wilcoxon W	3532.500	3004.500	1372.000	137.500	3621.000
Z	-.586	-4.190	-2.377	-2.479	-.047
Asymp. Sig. (2-tailed)	.558	.000	.017	.013	.962
Exact Sig. [2*(1-tailed Sig.)]				.012 <sup>a</sup>	

a. Not corrected for ties.

b. Grouping Variable: DST\_