AFRL-RH-WP-TR-2009-0021

Target Acquisition Involving Multiple Unmanned Air Vehicles: Interfaces for Small Unmanned Systems (ISUS) Program

Thomas R. Carretta
Michael J. Patzek
Lamar Warfield
Sarah E. Spriggs
Allen J. Rowe
Airam Gonzalez-Garcia
Kristen K. Liggett
Warfighter Interface Division
Supervisory Control Interfaces Branch

March 2009

Final Report for April 2006 to February 2009

Approved for public release;
Distribution is unlimited.

Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Warfighter Interface Division
Supervisory Control Interfaces Branch
Wright-Patterson AFB OH 45433

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th ABW Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2009-0021 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

__//signed//_____
Kristen K. Liggett, Ph.D. Work Unit Manager
Supervisory Control Interfaces Branch

_//signed//_____
Daniel G. Goddard, Chief
Warfighter Interface Division
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

*Disseminated copies will show "//signature//" stamped or typed above the signature blocks.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Service, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
23-03-2009	Final	April 2006 – February 2009

4. TITLE AND SUBTITLE
Target Acquisition Involving Multiple Unmanned Air Vehicles: Interfaces for Small Unmanned Air Systems (ISUS) Program

5a. CONTRACT NUMBER
NA

5b. GRANT NUMBER
NA

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S
Thomas R. Carretta, Michael J. Patzek, Lamar Warfield, Sarah E. Spriggs, Allen J. Rowe, Airam Gonzalez-Garcia, & Kristen K. Liggett

5d. PROJECT NUMBER
7184

5e. TASK NUMBER
09

5f. WORK UNIT NUMBER
71840917

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
711th Human Performance Wing
Human Effectiveness Directorate
Warfighter Interface Division
Supervisory Control Interfaces Branch
Wright-Patterson AFB OH 45433-7501

10. SPONSOR/MONITOR'S ACRONYM(S)
711 HPW/RHCI

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RH-WP-TR-2009-0021

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; Distribution is unlimited.

13. SUPPLEMENTARY NOTES
88 ABW/PA cleared 03/12/09; 88ABW-09-0990.

14. ABSTRACT

The use of small unmanned aerial vehicles (UAVs) and micro air vehicles (MAVs) in military reconnaissance, surveillance, and target acquisition (RSTA) missions is becoming increasingly widespread. The Supervisory Control Interfaces Branch of the Air Force Research Laboratory is exploring interface design techniques to support multiple-UAV, single-operator employment concepts. These employment concepts require that a single operator monitor or manage multiple vehicles and or sensors which may increase visual/cognitive workload and degrade performance. This paper reviews four experiments performed for the Interfaces for Small Unmanned Systems (ISUS) Program involving target acquisition in the multiple-UAV context. The goals were to characterize and improve human performance. Experiment 1 examined unaided operator performance in a target acquisition task involving multiple MAV videos. Experiment 2 compared operator performance when target acquisition was aided and unaided for multiple simulated MAV videos. Experiment 3 examined the effect of display size on unaided target acquisition for multiple MAV videos. Experiment 4 compared target acquisition performance for unaided human operators with that of an automated cooperative controller in accomplishing a complex task involving the prosecution of ground based targets with Wide Area Search Munitions (WASMs). The findings and lessons learned are the basis for recommendations that could provide valuable insights into the size and configuration of the sensor display for performing target acquisition, especially those requiring simultaneous monitoring of multiple videos. The findings also may contribute to shaping employment concepts and technology requirements for future unmanned aerial systems.

15. SUBJECT TERMS
Target acquisition, reconnaissance/surveillance, micro air vehicles

16. SECURITY CLASSIFICATION OF: Unclassified			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Kristen K. Liggett
U	U	U	SAR	81	19b. TELEPONE NUMBER (*Include area code)*

This page intentionally left blank.

Table of Contents

FIGURES

TABLES

PREFACE

This report describes activities performed in support of the Air Force Research Laboratory Warfighter Interface Division, Supervisory Control Interfaces Branch (AFRL 711 HPW/RHCI) Unmanned Aerial Vehicle Interface Defense Technology Objective (UAV DTO), Work Unit 71840917.

This page intentionally left blank.

INTRODUCTION

Unmanned aerial vehicles (UAVs) are becoming more widespread in military operations, especially in performing reconnaissance and surveillance missions. The military role of UAVs has grown at an unprecedented rate. In 2005, tactical and theater level UAVs flew over 100,000 hours in support of Operation Iraqi Freedom and Operation Enduring Freedom. According to a January 2008 Associated Press article, the US Air Force more than doubled its use of drones between January and October 2007 while the number of unmanned flight hours for DoD systems soared to over 500,000 (USA Today, 2008). The explosion in UAV use has been spurred on by technological improvements enabling greater capability to be placed on smaller platforms. Several key UAV mission areas leverage their long endurance and surveillance capabilities (Office of the Secretary of Defense, 2001). UAV mission areas are expanding, along with capabilities to afford more responsive, persistent, and integrated operations. As more unmanned aerial systems are incorporated into everyday military operations and as their roles become more demanding, efforts are underway to advance operator interface technology to improve human performance, system capability and overall mission effectiveness.

Science and technology efforts are focused on enhancing user situation awareness and control performance, with a goal to reduce the number of crew members required to operate unmanned systems (Barbato, 2000). Ultimately, the goal is to enable one operator to manage multiple unmanned systems for very complex missions (Clough, 2002; Ruff, Calhoun, Draper, Fontejon, & Guilfoos, 2004; Scerri, Liao, Lai, Syeara, Xu, & Lewis, 2005; Walker, 2005). Many factors influence if and how this can be accomplished, such as the nature of the mission(s), the class and on-board capabilities of the UAV(s), the operator station (i.e., human-system interface), and the operators' background and skills. The requirement to monitor several unmanned systems simultaneously will likely increase visual search requirements and mental workload and adversely impact operator situational awareness and performance (Alexander, Nygren, & Vidulich, 2000; Dixon & Wickens, 2003; Wickens, Dixon, & Chang, 2003). Alexander et.al. (2000) examined the relations between mental workload and situational awareness in a simulated combat task. The task scenario consisted of four phases designed to

influence pilot mental workload and situational awareness. Results indicated that as mental workload increased operator situational awareness decreased.

The Supervisory Control Interfaces Branch of the Air Force Research Laboratory is exploring multi-UAV, single-operator concepts for conducting reconnaissance, surveillance, and target acquisition (RSTA) missions. The control of multiple systems allows for force multiplication, where one system conceivably can accomplish the job of several systems. However, a human factors consequence of this is that one operator may need to manage more information than did operators of legacy systems. Therefore, it is important to test and evaluate information portrayal in this new context to ensure that human-system interface technology evolves, along with system capabilities, to meet the warfighters' needs.

A key responsibility of the operator is to detect, identify, and possibly monitor ground objects and targets using video or still imagery sent to the operator by the vehicles' sensors. The operator may need to continuously monitor sensor returns (e.g., UAV or micro air vehicle [MAV] videos) to acquire targets of interest. This task can be visually and cognitively demanding given the scanning, attentional, perceptual, and memory requirements to focus on each video and monitor the changing scenes.

This report reviews four experiments conducted by the Interfaces for Small Unmanned Systems (ISUS) program involving target detection in the multiple UAV context. The goals were to characterize and improve human performance. Experiment 1 examined unaided operator performance in a target acquisition task involving multiple MAV videos. Experiment 2 compared operator performance when target acquisition was aided and unaided for multiple simulated MAV videos. Experiment 3 examined the effect of display size on unaided target acquisition for multiple MAV videos. Experiment 4 compared target acquisition performance for unaided human operators with that of an automated cooperative controller in accomplishing a complex task involving the prosecution of ground based targets with Wide Area Search Munitions (WASMs) (Scerri et.al., 2005).

EXPEREIMENT 1: TARGET ACQUISITION INVOLVING MULTIPLE MAV VIDEOS

This experiment had two parts. Part one investigated unaided operator target acquisition performance using single and multiple MAV video presentations. Part two examined subjective ratings of image quality/interpretability for still and video imagery.

Method

Participants

Participants were 24 civilian and military full-time employees stationed at Wright-Patterson AFB, OH. The sample consisted of 22 men and two women who ranged in age from 20 to 49 years with a mean age of 34.3 years. All participants reported being in good to excellent health with vision correctable to 20/20, normal color vision, and normal peripheral vision. Most of them had previous experience with simulators (54.2percent) and video games (75.0percent). No compensation was offered in exchange for participation in the study.

Materials

MAV Videos

The videos used for this experiment were recorded from a forward-looking color camera at a 45 degree depression mounted in the nose of a MAV. The camera had a resolution of 380 lines, 76 degree field of view, and a 2.4 GHz wireless data link for video with a 900 MHz 2-way modem. The MAV flew at an altitude of 150 feet with an airspeed of 22 knots. Several videos of approximately 15-20 minutes in length were edited to create 28 one-minute clips for use as test material. The terrain in the videos was open with a fairly flat surface, roads, and clusters of trees and shrubs. Several man-made objects appeared in the videos including a bridge, buildings, tanks and other vehicles, a scud launcher, a surface-to-air missile site, and an airplane runway.

Data Collection Instruments

Both objective and subjective measures were collected. The questionnaires (Warfield, Carretta, Patzek, & Gonzalez-Garcia, 2007) are described below.

Objective measures of target detection performance. Two measures of target detection performance were collected: number of hits and number of false alarms.

Demographic data questionnaire. This questionnaire (Appendix A) was used to collect information to characterize the sample and assist in interpretation of participants' performance on the target detection task. Items elicited information about participants' gender, age, general health, wellbeing, previous experience with simulator-type environments and with video games, and whether they had vision correctable to 20/20 acuity and normal peripheral and color vision.

Confidence ratings. Whenever participants detected a target, they were instructed to indicate the level of confidence in their target detection decision. Confidence ratings were made on a five-point Likert rating scale (1 - not at all confident, 2 - slightly confident, 3 - moderately confident, 4 - fairly confident, 5 – very confident). Participants also were asked to indicate their confidence after viewing videos in which they determined no targets existed (i.e., how confident were they in their decision not to designate anything as a target).

Civil National Imagery Interpretability Rating Scale (NIIRS). The NIIRS (Appendix B) is a task-based scale used to rate the quality and interpretability of imagery acquired from imaging systems (Irvine, 1997; Riehl & Maver, 1996). The NIIRS originated in the intelligence community and is the standard used by collection managers, imagery analysts, and sensor designers (Leachtenhauer, 1996; Maver, Erdman, & Riehl, 1995). It provides a common framework or standard for describing the information potential or interpretability of imagery. The NIIRS has 10 ordered levels (0 to 9) where each level is described by several interpretation tasks. The example tasks indicate the information detail that can be derived from an image of a given interpretability level. Several versions of the NIIRS are available including Civil (non-military). Visible (military), Radar (Synthetic Aperture Radar), Infrared (IR), and Multi-spectral. The Civil NIIRS was used in this study instead of the Visible NIIRS as the participants were not expected to be familiar with the military imagery used as anchors in the Visible NIIRS, but could be expected to be familiar with the non-military imagery used as anchors in the Civil NIIRS. Appendix B contains the NIIRS questionnaire.

Image Quality Rating Scale (IQRS). The IQRS (Appendix C) is a five-point Likert rating scale (1 – very poor, 2 – poor, 3 – fair, 4 – good, 5 – very good) that was created for this study.

Unlike the NIIRS, which uses task-based descriptions to define its levels of image interpretability, the IQRS merely used simple qualitative anchors. Irvine and his colleagues (Irvine, Fenimore, Cannon, Roberts, Israel, Simon, Watts, Miller, Brennan, Aviles, Tighe, & Behrens, 2005; Irvine, Fenimore, Cannon, Roberts, Israel, Simon, Watts, Miller, Brennan, Aviles, Tighe, Behrens, & Haverkemp, 2005) contend that the standard methodology used for NIIRS development for still imagery may not be applicable, or may require modification for application to motion imagery. They have proposed the use of a bipolar rating scale for comparing image quality/interpretability similar to the IQRS used in the current study (Corriveau, Gojmeiac, Hughes, & Stelmach, 1999). Appendix C contains the IQRS questionnaire.

Post-test questionnaire. This questionnaire (Appendix D) was used to elicit information regarding participants' assessment of the video imagery used in the study. The video imagery was rated in terms of quality, clarity, contrast, and resolution on a five-point scale (1 - poor, 2 - fair, 3 - good, 4 - very good, 5 – excellent). Participants also were given the opportunity to provide comments regarding video quality and other factors that affected their ability to detect targets.

Equipment

Two side-by-side 20-inch monitors were used to display still target images and videos. The still targets were displayed on a 20-inch color cathode ray tube (CRT) monitor and the videos on a 20-inch liquid crystal display (LCD) monitor. The 20-inch LCD monitor had a resolution of 1600 x 1200 pixels and was divided into quadrants, each quadrant displaying a different MAV video.

Procedures

The study required about two hours and was completed during a single testing session with short breaks as needed. Initial procedures included an overview of the study, informed consent, demographic data collection, and familiarization with the equipment. Participants then

completed the target detection and image quality rating tasks. Data collection began with completion of the demographic data questionnaire.

Part One: Target Detection

Participants were required to locate targets that appeared in one-minute videos. The number of videos monitored was varied to either be one, two, or four. Two side-by-side computer screens were used. The right screen (Figure 1) displayed the videos while the left screen (Figure 2) displayed six photographs that closely resembled the potential targets in the video. The primary task was to locate targets in the MAV videos. Participants were instructed that when they observed a target embedded in a video to use the mouse to place the cursor anywhere on the video containing the target and click the mouse. In addition to the primary task of locating targets, participants were given a secondary task. While searching for targets, participants were required to count the man-made objects (e.g., vehicles, buildings, etc.) that appeared in the videos. The measures of interest were target detection accuracy (i.e., hits and false alarms) and confidence in target/non-target decisions.

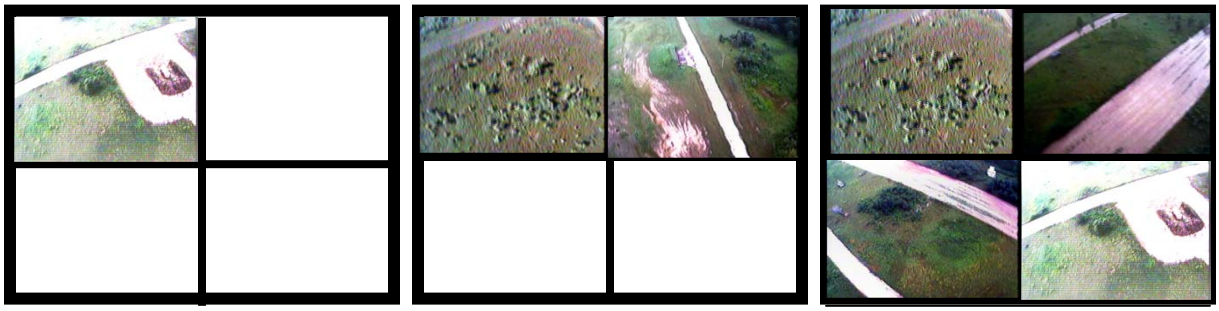


Figure 1. Example of one, two, and four screen display formats. When only one video was presented it always appeared in the upper left quadrant. When two videos were presented they always appeared in the upper left and right quadrants.

Each level occurred eight times, thus each participant viewed a total of 24 videos (3 levels of video monitoring x 8 replications = 24 presentations). The assignment of the video clips to the three levels of video conditions was randomized as was the ordering of the 24 presentations. Unbeknownst to the participants, half of the 24 presentations contained an actual target and half did not. Each target type was presented twice and target presentation was counter-balanced across the video conditions.

Figure 2. Ground-based images of targets used in Part 1.

Immediately following each video presentation session, participants were asked to indicate their level of confidence in their target/non-target decisions. If participants indicated they observed more than one target for a one-minute item, separate confidence ratings were made for each observation. If no targets were observed for an item, participants rated their confidence that there were no targets for that item. Participants were then asked how many man-made objects they counted. This procedure was repeated until all 24 video presentations were completed.

Part Two: Image Quality/Interpretability Evaluation

In part two, participants rated the image quality/interpretability of still photographs and short (15 second) video clips that contained various targets of interest. Image quality ratings were made using the Civil NIIRS and an image quality rating scale (IQRS). Participants studied the still image and video clip each for 15 seconds. Consistent with Irvine et.al. (2005a, 2005b), for each still/video pair the still image was viewed and evaluated first followed by the video. The purpose of these comparisons was to examine participants' ratings of the relative quality/interpretability of still versus video imagery for target detection and identification. After

completion of part two, participants completed a questionnaire regarding the clarity, contrast, resolution, and interpretability of the videos.

Analyses

As the data were not expected to be normally distributed, nonparametric tests were used to examine differences in performance across conditions (e.g., number of videos monitored). The Friedman test is a nonparametric test that compares three or more related groups (e.g., participants' performance while monitoring one, two, and four video conditions). First, the values for each participant are ranked from low to high. Performance for each participant is ranked separately. Next, the ranks for each condition (number of videos monitored) are summed. If the sums of the ranks for each condition are very different, the probability level will be small. The Friedman test was used for all comparisons involving three related groups.

Correlations were computed to examine relations between demographic variables (age, video game experience, and simulator experience) and measures of performance in the target detection task. Both Pearson (based on scores) and Spearman (based on ranks) correlations were computed.

The Wilcoxin matched pairs test was used for analyses involving only two related groups (e.g., Civil NIIRS ratings for still versus video imagery, IQRS ratings for still versus video imagery). Like the Freeman test, it is a nonparametric test used to examine distributional differences in performance for related groups. The Wilcoxin test calculates the difference between each set of pairs, ranks the differences (positive or negative), and analyzes that list of differences. If the two sums of ranks are very different, the probability level will be small. All comparisons were made using a 0.05 Type I error rate.

Experiment 1: Results and Discussion

Target Detection

The overall hit rate was about 75percent and was not affected by the number of videos (Figure 3). However, the number of false alarms increased significantly as the number of videos increased from one (1.37) to two (2.16) to four (4.83) (χ^2 (2) = 22.52, $p < 0.01$)(Figure 4).

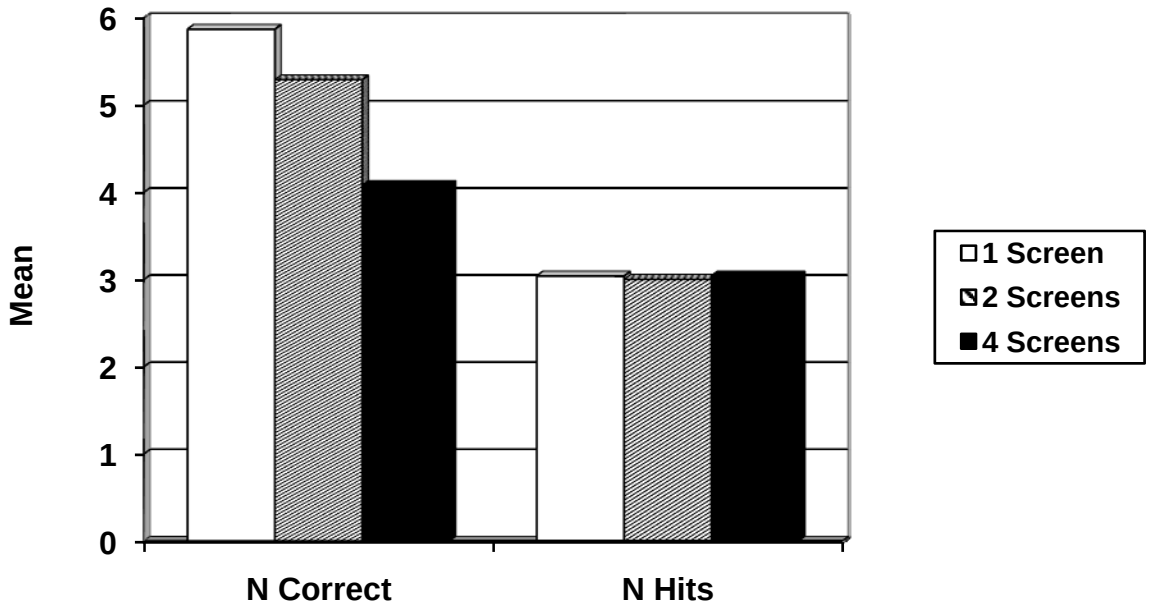


Figure 3. Target Detection Task: Objective measures of performance by number of videos monitored simultaneously. Measures included number correct and number of hits.

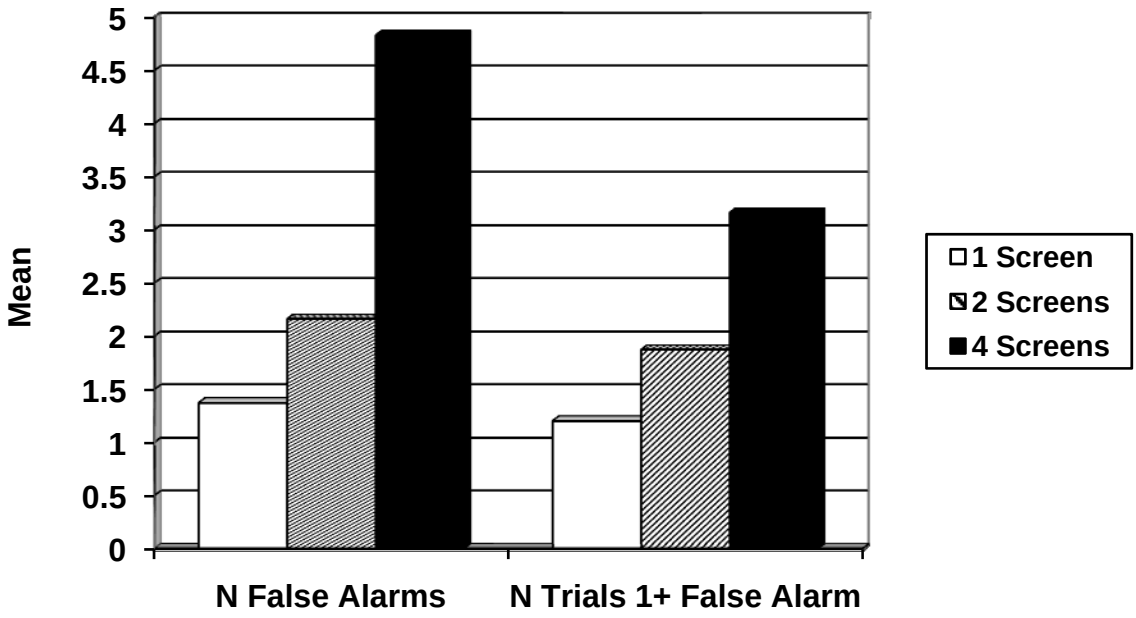


Figure 4. Target Detection Task: Objective measures of performance by number of videos monitored simultaneously. Measures included number of false alarms (N False Alarms) and number of trials with one or more false alarms (N Trials 1+ FA).

The correlation between target exposure time and detection rate was 0.70. When target size also was included as a predictor of detection rate, the multiple R increased to 0.76. Self-confidence in target detection decisions decreased significantly as the number of video presentations increased from one (3.80) to two (3.51) to four (3.41) (χ^2 (2) = 17.61, $p < 0.01$). It was surprising that the target detection rate was not affected by the number of videos monitored. This result may have been a function of task characteristics including few targets/distracters and short length of the videos (1 minute).

Image Quality/Interpretability

Ratings of image quality/interpretability were significantly higher for video versus still imagery for both the NIIRS (still = 3.89, video = 4.51; T = -4.17, $p < 0.01$) and IQRS (still = 2.43, video = 2.96; T = -4.03, $p < 0.01$) (Table 1). Although the magnitude of the difference in image quality/interpretability was not large, participants consistently rated the video imagery higher in quality/interpretability. Twenty-two of twenty-four participants (91.7 percent) rated the video imagery higher than the still imagery for quality/interpretability on both rating scales (Figure 5). Preference for videos versus still imagery may have been due to emergent properties of video (i.e., contextual cues). Also, videos provide a more natural way of interpreting imagery. The preference for video imagery is supported by Irvine et.al. (2005b) who also reported that target motion had a significant effect on perceived image quality. Video imagery clips in which the targets were moving were consistently rated higher than video clips with stationary targets. The preference for moving targets may be due to their increased salience (Irvine et.al., 2005b).

Lessons Learned

It was speculated that the reason target detection performance was not affected by the number of videos may have been due to task characteristics such as the low density of targets and non-targets in the videos and the short video length. To examine this, Experiment 2 used an environment rich in both targets and non-targets and longer videos (5 minutes). Also, a subjective measure of workload was used to assess perceived task demands. Finally, a target cueing capability was added to examine its effect on target acquisition performance.

Table 1. Image Quality/Interpretability Ratings for Still and Video Imagery

Score	Mean	SD	Negative Ranks	Positive Ranks	Tie Ranks	*T*
Civil NIIRS						
Still Images	3.89	0.88	22	1	1	-4.17**
Video Images	4.51	0.82				
Image Quality Rating Scale						
Still Images	2.43	0.49	22	2	0	-4.03**
Video Images	2.96	0.43				

Notes. Positive ranks occurred when the still images received a higher rating than did the video images. Negative ranks occurred when the still images received a lower rating than did the video images. Tie ranks occurred when the still and video images received the same rating.
N = 24; ** p < .01

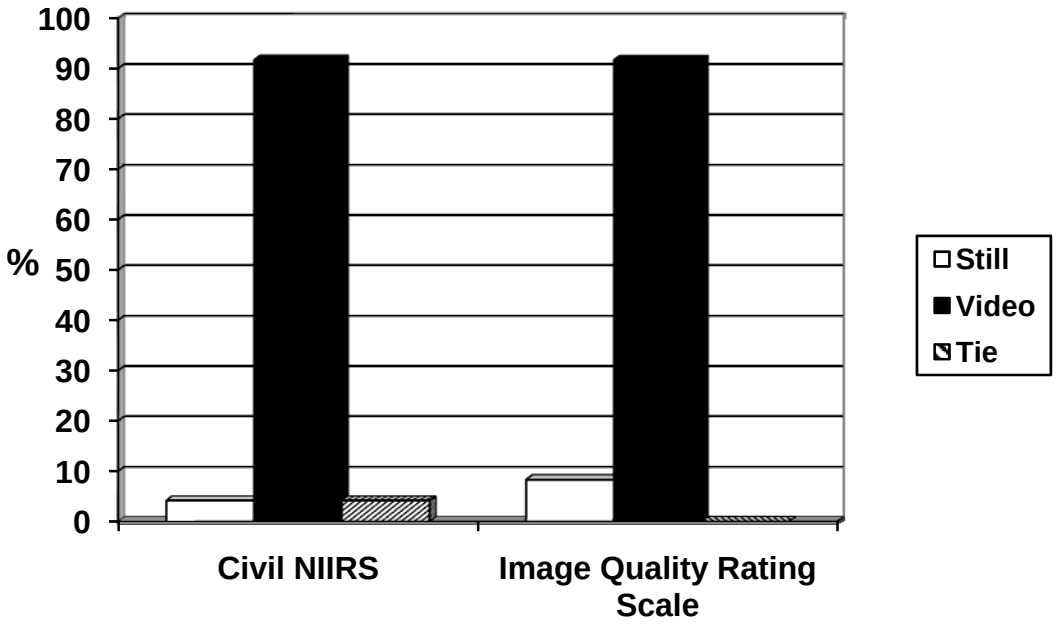


Figure 5. Participants' preferences regarding still versus video imagery.

EXPERIMENT 2: AIDED VS. UNAIDED TARGET ACQUISITION INVOLVING MULTIPLE SIMULATED MAV VIDEOS

Method

Experiment 2 examined operator performance and subjective workload for a target detection task similar to that performed by an UAV sensor operator. A computer-generated, synthetic environment was used to simulate MAV video in a scenario that replicated potential urban reconnaissance operations for MAVs. The impact of an automated target cueing (ATC) capability on target detection accuracy and workload was evaluated. Two levels of ATC capability were evaluated, in addition to a baseline of no cueing.

Participants

Eighteen full-time civilian and military employees stationed at Wright-Patterson Air Force Base, OH participated in this study. This sample consisted of 15 men and 3 women who ranged in age from 24 to 56 years with a mean of 31.5 years. All participants reported being in good to excellent health and having vision correctable to 20/20, normal color vision, and normal peripheral vision. Some participants indicated that they had previous video game (28percent) and UAV (33percent) experience. Participation was voluntary and no compensation was offered in exchange for participation in this study.

Materials

Simulated MAV Videos

Each simulated MAV video was five minutes long. The MAVs followed roads in the simulated environment. Static (non-moving) vehicles were positioned on each side of the road. Two-hundred fifty vehicles were placed beneath the five minute flight path of each MAV. Several models of civilian vehicles appeared in a variety of colors. Ten of the 250 vehicles in each video were targets; the remaining 240 vehicles were “distracters.” The target was a sport-utility vehicle (SUV) with its tailgate swung open (Figure 6). The target faced in each direction of traffic. Automatic target cueing level 1 (90 percent hits, 0.83 percent false alarms) identified nine of the ten targets and indicated two of the 240 distracters as targets (false alarms).

Automatic target cueing level 2 (60 percent hits, 2.08 percent false alarms) identified six of the ten targets and identified five of the 240 distracters as targets (false alarms). The cueing technique in this study presented the corners of a square around targets and false alarms. These corners appeared in magenta.

Figure 6. The target vehicle shown in white. This is an example of the ground-truth images participants were shown during familiarization.

Data Collection Instruments

Both objective and subjective measures of performance were gathered. Objective measures included the number of hits and false alarms that participants had in the target detection task during each trial. Subjective ratings of workload were recorded following each five-minute test session. A demographic data questionnaire was used to characterize the sample and a post-test questionnaire elicited comments about experimental procedures. The questionnaires are described below and are available in Petkosek, Carretta, Patzek, and Stoor (2007).

Objective measures of target detection performance. Two measures of target detection performance were collected: number of hits and number of false alarms.

Demographic data questionnaire. This questionnaire (Appendix E) was used to collect information to characterize the sample and assist in interpretation of participants' performance on the target detection task. Items elicited information about participants' gender, age, general health, wellbeing, previous experience with simulator-type environments and with video games, and whether they had vision correctable to 20/20 acuity and normal peripheral and color vision.

NASA Task Load Index (NASA-TLX). The NASA TLX (Hart & Staveland, 1988) is a subjective workload assessment tool (Appendix F). A multidimensional weighting procedure is used to derive an overall workload score based on weighted averages of ratings on 6 subscales: mental, physical, temporal, effort, performance, and frustration.

Post-test questionnaire. This questionnaire (Appendix G) was used to elicit information regarding participants' assessment of their performance. Participants rated the difficulty of the task, the effectiveness of the cueing, and their confidence in their target identification ability.

Equipment

The experiment was conducted in the Aerospace Vehicle Technology Assessment and Simulation (AVTAS) facility of the Air Vehicles Directorate at Wright-Patterson Air Force Base, OH. Participants were seated in an adjustable chair at a table with its surface 28" above the floor. They used an optical, four-button Microsoft mouse with a scroll wheel to designate their target detection decisions. A laptop computer was placed nearby where participants entered survey responses.

The Urban Simulation Environment is a custom-built network that consists of six standard computers. One of the computers served as the user station. Another simulated the MAVs' flights. The remaining four computers generated the simulated video feeds coming from each of the four MAVs.

The operator station consisted of an in-house OpenGL program that displayed the imagery from the MAV computers. The imagery was sent over a reflective memory network with less than one millisecond latency.

The MAVs waypoint-following capabilities were based on an existing system. The waypoints were laid out using a map of the area and flight paths were constructed to follow major roads. Care was taken in building non-overlapping MAV flight paths. Flight paths were fed to the flight control computer for each MAV. Each trial was a real-time flight. There were no issues with repeatability, because the flight models were deterministic. The flight models were

synchronized to real time using an in-house executive that also controlled the programs that interfaced between the flight models and the MAV computers.

The MAV computers interfaced with the flight models using the Common Image Generator Interface. This open standard provided the ability to control various aspects of the scene including model locations, time of day, and video size. The models in the scene consisted of three-dimensional representations of common civilian automobiles in various colors. The time of day was set to 1200 hours for each run to eliminate variance in shadows. The video size was set to recreate the video feed coming from the MAVs with a resolution of 640 x 480 pixels. Full frames were sent over the network; there was no compression when the video was sent to the user station. The effects of interference, sun-blooming, and over exposure were not used in order to avoid further variances in test stimuli.

A laptop positioned on a different table was used to run the computer-based NASA-Task Load Index application.

Procedures

A mixed design was used. Participants were randomly assigned to one of the ATC conditions (no ATC, ATC level 1, or ATC level 2). Within each ATC condition, participants received all three video presentation levels (one, two, and four). Video presentation order was counterbalanced across participants.

The experimental procedure consisted of an introduction, and procedures, practice, and two test sessions. The introduction included an overview of the experiment, informed consent, demographic data collection, familiarization with the equipment, and an explanation of the NASA-TLX. Participants then completed a practice session to familiarize themselves with the equipment, procedures, and target detection task.

Four target vehicles (two black and two white SUVs) were used throughout the study. There were two side views and two forward-looking angular views from the rear of the vehicle, all taken from ground level. Once participants indicated they were comfortable with the targets, the pictures were removed. The experimenter then demonstrated the search task for one minute.

The ATC also was demonstrated if the participant was in an ATC group. Next, participants completed the demographic survey and were trained with the NASA-TLX.

As previously noted, each participant completed all three video presentation levels within one of three ATC conditions. The first session was practice and consisted of one trial for each of the video presentation levels. Sessions two and three were scored and also consisted of one trial for each of the video presentation levels. The NASA-TLX was administered following each trial. A mandatory five minute break was taken after the practice session. This was the only break in the experiment, unless participants requested that one be taken. After completion of the test sessions, participants completed a post-experiment questionnaire.

Analyses

All statistical tests used a 0.05 Type I error rate and were one-tailed. One-tailed tests were used because of results from prior research that led to directional expectations about how the experimental manipulations should affect performance. Specifically, it was expected that as the number of video presentations increased, target detection performance would decrease and subjective workload level would increase. Also, it was expected that as level of ATC improved, target detection performance would increase and subjective workload level would decrease. Further, it was expected that target detection performance and subjective workload level would be subject to an interaction between number of videos and level of ATC. Level of ATC was expected to have a greater effect on performance as the number of videos increased.

Experiment 2: Results and Discussion

Target Detection

Overall target detection rate was significantly higher in this experiment (95 percent) than in Experiment 1 (75 percent), which used live videos. Contrary to Experiment 1 where the target detection rate was not affected by the number of videos, a within subjects contrast indicated that detection rate decreased significantly as the number of videos monitored went from one (97.50 percent) to two (95.97 percent) to four (90.76 percent) ($F(1, 15) = 29.90$, $p < .01$). Also contrary to Experiment 1, the number of false alarms was very low and was not related to number of

videos or to ATC level. Surprisingly, target detection rate was not affected by ATC level (no ATC: 95.20 percent; Low ATC: 94.58 percent; High ATC: 94.44 percent) (Table 2; Figure 7).

Table 2. Means and Standard Deviations for Target Detection Performance by Number of Videos and Level of Cueing

N Videos/	1 Video		2 Videos		4 Videos		Combined	
Cueing	Mean	SD	Mean	SD	Mean	SD	Mean	SD
No Cueing	97.50	6.12	96.66	3.02	91.45	3.20	95.20	3.47
Low Cueing	97.50	4.18	96.66	2.04	89.58	6.59	94.58	3.52
High Cueing	97.50	2.73	94.58	1.88	91.25	4.67	94.44	1.94
Combined	97.50	4.28	95.97	2.44	90.76	4.79	94.74	2.90

Note. The number of videos was a within-subject variable, while level of cueing was a between-subjects variable. As a result, the sample size for each level of cueing was 6 and the total sample size was 18.

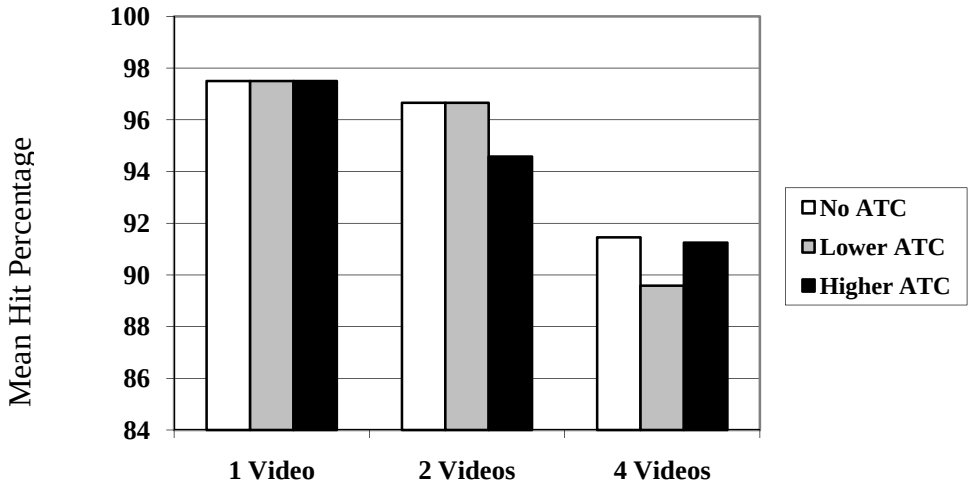


Figure 7. Mean hit percentage by number of videos and ATC capability.

Subjective Workload

A within-subjects contrast indicated that total workload increased as the number of videos monitored increased from one (10.77) to two (22.61) to four (56.27) ($F(1, 15) = 140.20$, p

< 0.01) (see Table 3; and Figure 8). However, total workload was not affected by ATC level. The trend for total workload also occurred for the six NASA-TLX subscales (mental, physical, temporal, effort, performance, and frustration). That is, level of subjective workload increased as the number of videos increased while ATC level had no effect on subjective workload ratings.

Table 3: Means and Standard Deviations for Total Subjective Workload by Number of Videos and Level of Cueing

N Videos/	1 Video		2 Videos		4 Videos		Combined	
Cueing	Mean	SD	Mean	SD	Mean	SD	Mean	SD
No Cueing	7.58	3.45	18.00	7.25	54.91	19.31	26.83	7.94
Low Cueing	8.25	4.55	20.25	15.85	59.66	19.00	29.38	12.57
High Cueing	16.50	13.52	29.58	17.75	54.25	11.45	33.44	13.40
Combined	10.77	8.98	22.61	14.44	56.27	16.14	29.88	11.21

Note. The number of videos was a within-subjects variable, while ATC capability was a between-subjects variable. As a result the sample size for each level of cueing was 6 and the total sample size was 18.

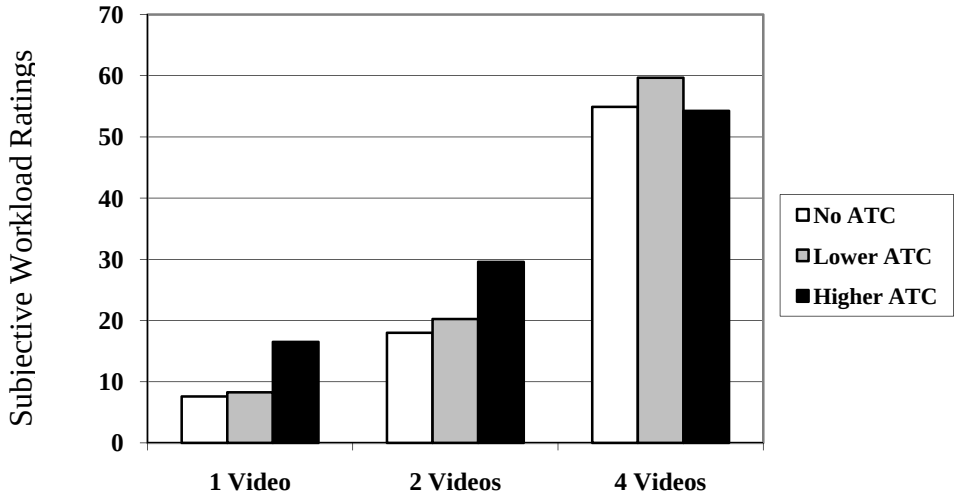


Figure 8. Mean total subjective workload rating by number of videos and ATC capability.

Lessons Learned

The lack of an effect of ATC on target acquisition was consistent with findings from Janson, See, Riegler, Davis, and Kuperman (1998) and See, Davis, and Kupperman (1997) who examined automatic target cueing and target localization performance with forward-looking infrared (FLIR) and synthetic aperture radar (SAR), respectively. Both studies found that ATC increased operators' confidence in their decisions, but did not affect localization accuracy or perpetual sensitivity compared to not using the ATC. These studies both concluded that ATC would be most beneficial when the viewing conditions exceeded an operator's ability to effectively locate targets. See et.al. (1997) speculated that a poor ATC capability could worsen performance compared to what would be achieved without it.

There are at least two potential reasons for the lack of ATC impact on target detection effectiveness in the current study. First, target detection level was very high throughout the experiment, even in the unaided (no ATC) condition where the average detection rate was 95.2 percent across the three video presentation conditions As in Janson et.al. (1998) and See et.al. (1997) participants were able to effectively locate targets without the ATC. More thorough pre-testing should have been done to identify an appropriate target detection difficulty level. Second, a visually-based ATC may have been ineffective because it did not reduce operator requirements to conduct a visual search. The way the ATC was implemented (using visual cues) required participants to scan the multiple screens which they were required to do anyway, thus the target detection strategy was probably not significantly different in the no ATC and ATC conditions. Future studies should use a different cueing method (e.g., auditory, haptic) to direct the operator's attention to potential targets and reduce the visual requirement to scan all screens continuously.

EXPERIMENT 3: EFFECT OF VIDEO DISPLAY SIZE ON UNAIDED TARGET ACQUISITION INVOLVING MULTIPLE MAV VIDEOS

Method

The objective of Experiment 3 was to compare target acquisition performance using different video display sizes (i.e., 5” by 6 3/4” with 640 x 480 resolution versus 2 1/2” by 3 5/16” with 320 x 240 resolution) for both single and multiple video presentations. Potential advantages of smaller video displays include less area to scan, a perception of less video jitter, and more flexibility in display configuration and design. A potential drawback of smaller video displays is that the targets of interest will be relatively smaller on the screen and be represented by fewer display pixels, which may adversely affect target detection performance. Unlike the previous experiments in which participants monitored one, two, or four videos, in the current experiment, participants monitored either one small, one large, four small, or four large videos. Figure 9 shows examples of video layouts for the one and four video presentation configurations using small and large displays.

Figure 9. Example of video layouts.

Participants

Participants were 16 civilian and military full-time employees stationed at Wright-Patterson AFB, OH. The sample consisted of 15 men (93.8 percent) and one woman (6.3 percent). Participants ranged in age from 24 to 51 years with a mean age of 32.8 years. All participants reported being in good to excellent health and having visual acuity corrected to 20/20, normal color vision, and normal peripheral vision. Most participants indicated they had previous experience with simulators (63 percent) and video games (56 percent). Participation was voluntary and participants could withdraw from the study at any time without penalty. No compensation was offered in exchange for participation in this study.

Materials

MAV Videos

The videos used for this experiment were recorded from a forward-looking color camera at a 45 degree depression mounted in the nose of a MAV. The camera had a resolution of 640 X 480 lines, 30 degree field of view, and a 2.4 GHz downlink (wireless data link) for video with a 900 MHz 2-way modem. The video was streamed at 30 frames per second. The MAV flew at approximately 175 feet.al.titude above the ground with an airspeed of approximately 22 knots. Several videos of about 15-20 minutes in length were edited to create 12 five-minute clips for use as test material and five one-minute clips for pre-test training materials. Several buildings, roads, and vehicles were dispersed over the setting.

Data Collection Instruments

Both objective (hits, false alarms) and subjective (workload) measures of performance were included. As with the previous experiments, a demographic data questionnaire was used to characterize the sample and a post-test questionnaire elicited comments about experimental procedures. The questionnaires are described below and are available in Plantz, Warfield, Carretta, Gonzalez-Garcia, and Patzek (2008).

Objective measures of target detection performance. Two measures of target detection performance were collected: number of hits and number of false alarms. Number of hits was

converted to a percentage as there were different video arrangements and number of targets in the one and four video configurations.

Demographic data questionnaire. This questionnaire (Appendix A) is the same as that used in Experiments 1.

Confidence ratings. Whenever participants detected a target, they were instructed to verbally state the type of target (shelter, SUV, truck, and van) and the level of confidence in their target detection decision. Confidence ratings were made on a five-point Likert rating scale (1 - not at all confident, 2 - slightly confident, 3 - moderately confident, 4 - fairly confident, 5 – very confident).

NASA-TLX. This measure is the same as that described in Experiment 2 (Appendix F).

Post-test questionnaire. This questionnaire (Appendix H) elicited information regarding participants' assessment of the video imagery used in the study. After completion of the target detection task, participants rated the quality, clarity, contrast, resolution, and interpretability of the video imagery used in the study. Video quality, clarity, contrast, and resolution were measured using a five-point scale (1 – poor, 2 – fair, 3 – good, 4 – very good, 5 – excellent). Interpretability was measured as a dichotomous variable (1 – yes, 0 – no). Participants also rated image quality/interpretability separately for the one small, one large, four small, and four large video display conditions using the five-point scale. Finally, participants were given the opportunity to provide comments regarding video quality and other factors that affected their ability to detect targets.

Equipment

Two side by side 24-inch widescreen LCD monitors were used to display still images of the targets and the videos (Figure 10). The still images of the target were provided to aid the participants during target acquisition (Figure 11). Both monitors had a resolution of 1920 x 1200 pixels. The still targets were displayed on the left monitor and the videos were displayed on the right monitor.

Figure 10. Equipment configuration.

Whenever participants detected a target, they were instructed to verbally state the type of target and the level of confidence in their target detection decision. Participants' vocal responses were recorded using a Plantronics DSP 500 headset with a microphone.

A laptop positioned on a different table was used to run the computer-based NASA-Task Load Index application.

Procedures

Experiment 3 was conducted in the Crew Systems Integration Laboratory at Wright-Patterson AFB, OH. It required about two hours and was completed during a single testing session with a short break in the middle. Initial procedures included an overview of the experiment, informed consent, demographic data collection, familiarization with the equipment, and an explanation of the NASA-TLX. Participants then completed a practice session to

familiarize themselves with the equipment, procedures, and target detection task. The NASA-TLX was completed following each trial.

The practice trials used a different target set and video footage than did the test trials. This was done so participants could become familiar with the target detection task procedures, but not with the test stimuli. The practice trials were each one-minute in length and occurred in the fixed order of one small video, one large video, four small videos, and four large videos. The test trials were randomized across participants and occurred in a counterbalanced order that took into account video display size (large vs. small) and number of videos (1 vs. 4).

Figure 11. Ground-based pictures of targets.

For each practice and test trial, participants were required to locate, designate, and identify targets that appeared in videos. Participants were instructed that when they observed a target embedded in a video to use the mouse to place the cursor on the target, as close to the

target as they could, and click the mouse. In addition, the participants were instructed to call out the name of the target with their confidence rating within two seconds after they clicked on a target. The confidence rating scale and its values were displayed on a card under the right monitor throughout data collection as a reference. The number of videos monitored simultaneously was varied to be either one or four. Video display size also was varied to be either 5 by 6 ¾" with 640 x 480 resolution or 2 ½ by 3 5/16” with 320 x 240 resolution.

Each test trial was five minutes in length. Immediately following each video combination, participants were asked to rate their workload level for that trial using the NASA-TLX. The procedure was repeated until all eight video presentations were completed. After viewing all of the videos, participants completed a questionnaire regarding image interpretability and quality.

Analyses

Although the study design crossed display size and number of video displays, the analyses focused on comparing performance within the number of videos conditions. That is, analyses examined performance for display size (small vs. large) within each number of video displays condition (one or four). The decision not to examine the interaction of display size and number of videos on performance was due to the experimental design in that different sets of videos were used in the one and four video display conditions. As a result, the causes for any differences in pattern of performance between the one and four video conditions could be confounded by the different video sets used.

This was an exploratory study as we had no expectations as to which display size would be more effective based on results from prior studies. A 0.05 Type I error rate was used with a non-directional (two-tailed) hypothesis. Related-samples t-tests were performed since participants were exposed to both display sizes. Objective measures of performance were hit accuracy (percent) and number of false alarms. Number of hits was converted to a percentage as there were different video arrangements and number of targets in the one and four video configurations. Subjective measures included overall workload, confidence in target detection decisions, and image interpretability ratings.

Experiment 3: Results and Discussion

Target Detection

The mean target detection percent was 77.96 percent for the one video condition and 68.31 percent for the four videos condition. Related samples t-tests indicated that within the number of videos conditions (one or four videos) there was no significant difference in target identification percent for the small and large video sizes. Within the one video presentation, the detection rates were 78.12 percent for the small and 77.81 percent for the large sizes ($t(15) = 0.14$, ns). Within the four videos presentation, the detection rate was 68.31 percent for both the small and large video conditions ($t(15) = 0.00$, ns) (see Table 4 and Figure 12).

Table 4. Target Identification Accuracy: Target Identification Percent by Number of Videos and Video Display Size Combination

Condition	Mean	SD	Min.	Max.	SD_D	t (15)
One Video						
Small	78.12	9.63	60.00	90.00	8.64	0.14
Large	77.81	6.57	70.00	90.00		
Four Videos						
Small	68.31	10.47	44.83	82.76	10.53	0.00
Large	68.31	9.27	51.72	79.41		

Note. A 2-tailed related samples t-test was used to compare the mean difference in target detection percent for small vs. large displays within number of videos. SD_D is the standard deviation for the related samples t-test. Degrees of freedom (df) equals N pairs – 1 = 15. N = 16; * $p < .05$ (2-tailed)

The mean number of false alarms was 1.40 for the one video condition and increased to 2.56 for the four videos condition. The related samples t-tests indicated that for the single video presentation, there was no significant difference in the number of false alarms for the two display sizes (small = 1.37; large = 1.43; $t(15) = -0.18$, ns). However, the number of false alarm was

significantly greater for the small display compared to the large display when four videos were viewed (small = 3.25. large = 1.87; t (15) = 2.58, $p < 0.05$) (see Table 5 and Figure 13).

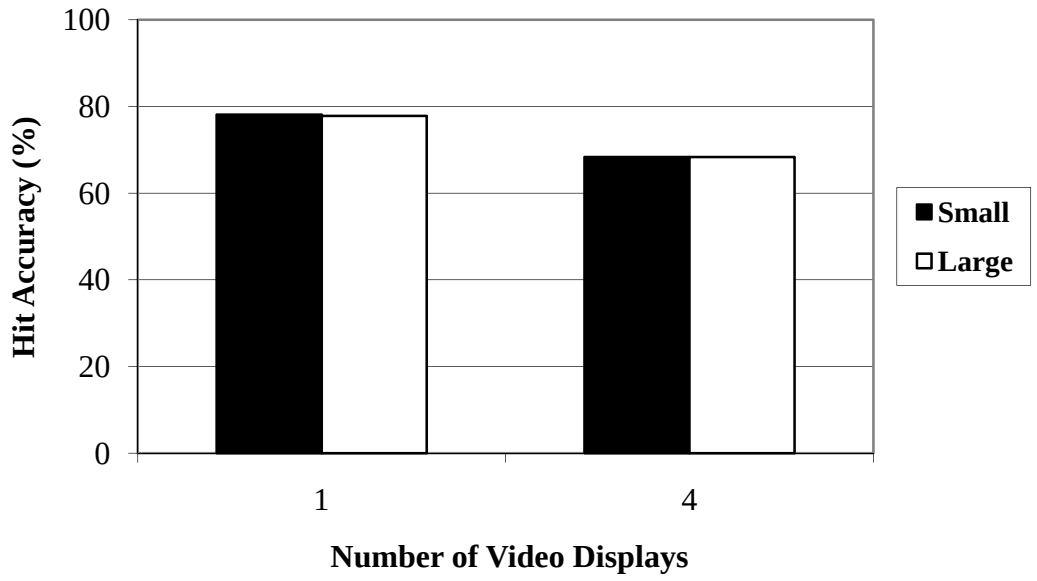


Figure 12. Hit accuracy (percent) by number of videos and video display size.

Table 5. Target Identification Accuracy: Number of False Alarms by Number of Videos and Video Display Size Combination

Condition	Mean	SD	Min.	Max.	SD_D	$t_{(15)}$
One Video						
Small	1.37	1.70`	0	5	1.34	-0.18
Large	1.43	1.86	0	5		
Four Videos						
Small	3.25	2.62	0	10	2.12	2.58*
Large	1.87	1.40	0	4		

Note. A 2-tailed related samples t-test was used to compare the mean difference in number of false alarms for small vs. large displays within number of videos. SD_D is the standard deviation for the related samples t-test. Degrees of freedom (df) equals N pairs – 1 = 15.
N = 16; *p < .05 (2-tailed)

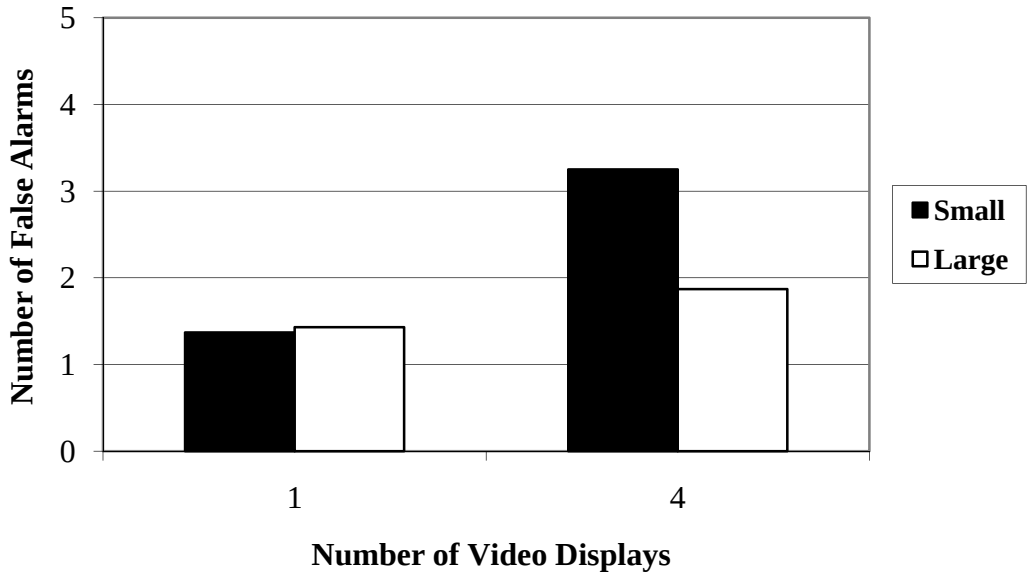


Figure 13. Number of false alarms by number of videos and video display size.

Confidence Ratings

There was little variability in confidence in target detection decisions across the number of videos by display size combinations. The related samples t-tests indicated there was no significant difference in average confidence rating for the two display sizes within each of the number of video display conditions. For the one video condition, the mean confidence ratings were 4.40 for the small and 4.33 for the large sizes (t(15) = 0.89, ns). For the four video condition, the mean confidence ratings were 4.25 (small) and 4.27 (large) (t(15) = -0.26, ns) (see Table 6 and Figure14).

Table 6. Confidence Rating by Number of Videos and Video Display Size Combination

Condition	Mean	SD	Min.	Max.	SD_D	t (15)
One Video						
Small	4.40	0.43`	3,47	5.00	0.32	0.89
Large	4.33	0.59	3.07	5.00		
Four Videos						
Small	4.25	0.53	3.06	5.00	0.29	-0.26
Large	4.27	0.53	2.93	5.00		

Note. A 2-tailed related samples t-test was used to compare the mean difference confidence ratings for small vs. large displays within number of videos. SD_D is the standard deviation for the related samples t-test. Degrees of freedom (df) equals N pairs – 1 = 15.

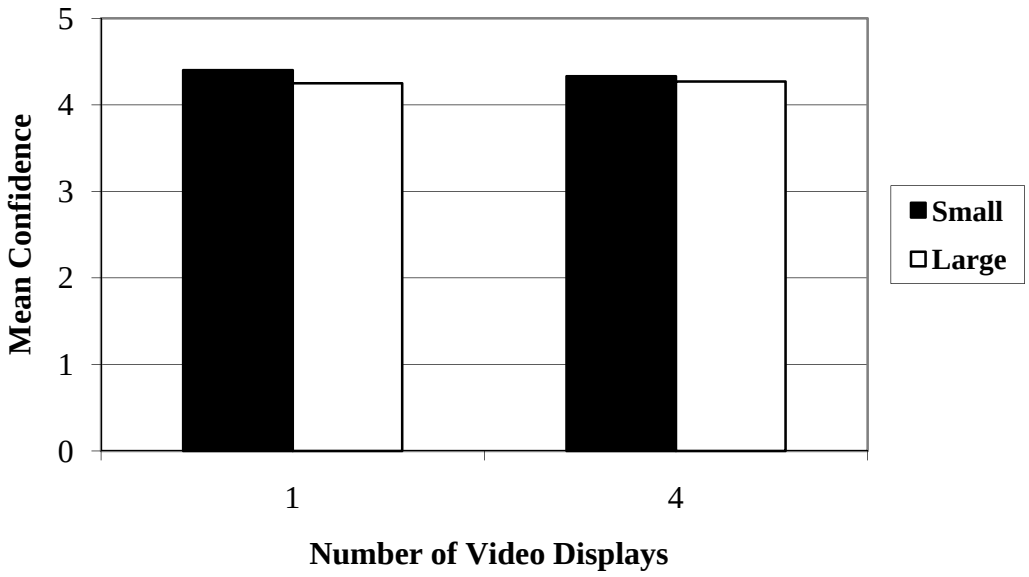


Figure 14. Mean confidence rating by number of videos and video display size.

Workload

The mean overall workload was 43.36 for the one video display condition and 54.77 for the four videos display condition. The related samples t-tests indicated that there was no significant difference in average overall workload for the two display sizes within each of the number of video display conditions. For the one display condition, the mean overall workload was 45.04 and 41.68 for the small and large displays (t(15) = 1.82, ns). For the four video condition, the means were 55.15 and 54.40 for the small and large displays (t(15) = -1.84, ns) (see Table 7 and Figure 15).

Table 7. Overall Workload by Number of Videos and Video Display Size Combination

Condition	Mean	SD	Min.	Max.	SD_D	t (15)
One Video						
Small	45.04	15.57`	10.83	68.50	7.35	1.82
Large	41.68	15.62	10.67	69.50		
Four Videos						
Small	55.15	20.47	15.50	87.83	17.98	-1.84
Large	54.40	20.97	12.67	88.00		

Note. A 2-tailed related samples t-test was used to compare the mean difference in overall workload for small vs. large displays within number of videos. SD_D is the standard deviation for the related samples t-test. Degrees of freedom (df) equals N pairs – 1 = 15.

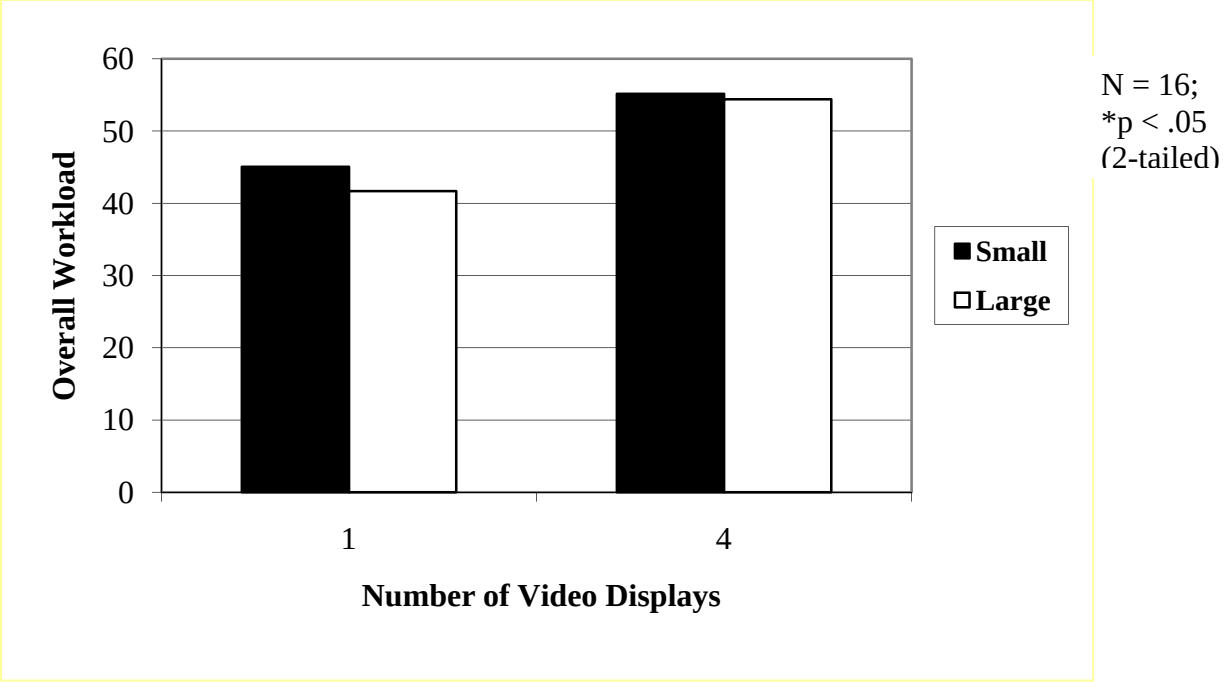


Figure 15. Overall workload by number of videos and video display size.

Video Quality and Interpretability

After completion of the target detection task, participants rated the quality, clarity, contrast, resolution, and interpretability of the video imagery used in the experiment. The related samples t-tests comparing the display sizes within the number of displays conditions indicated that image quality/interpretability rating was significantly lower for the small display relative to the large display in the one video presentation condition. Although the direction of the difference was the same for the four videos condition, the difference was not statistically significant (see Table 8 and Figure 16).

Table 8. Image Characteristics: Image Quality/Interpretability Rating by Number of Videos and Video Display Size Combination

Condition	Mean	SD	Min.	Max.	SD_D	t (15)
One Video						
Small	3.50	0.73`	2	5	0.44	-8.47**
Large	4.44	0.62	3	5		
Four Videos						
Small	3.00	0.81	2	4	0.96	-1.81
Large	3.44	0.62	2	4		

Note. A 2-tailed related samples t-test was used to compare the mean difference in image interpretability ratings for small vs. large displays within number of videos. SD_D is the standard deviation for the related samples t-test. Degrees of freedom (df) equals N pairs – 1 = 15.
N = 16; *p < .05; **p < .01 (2-tailed)

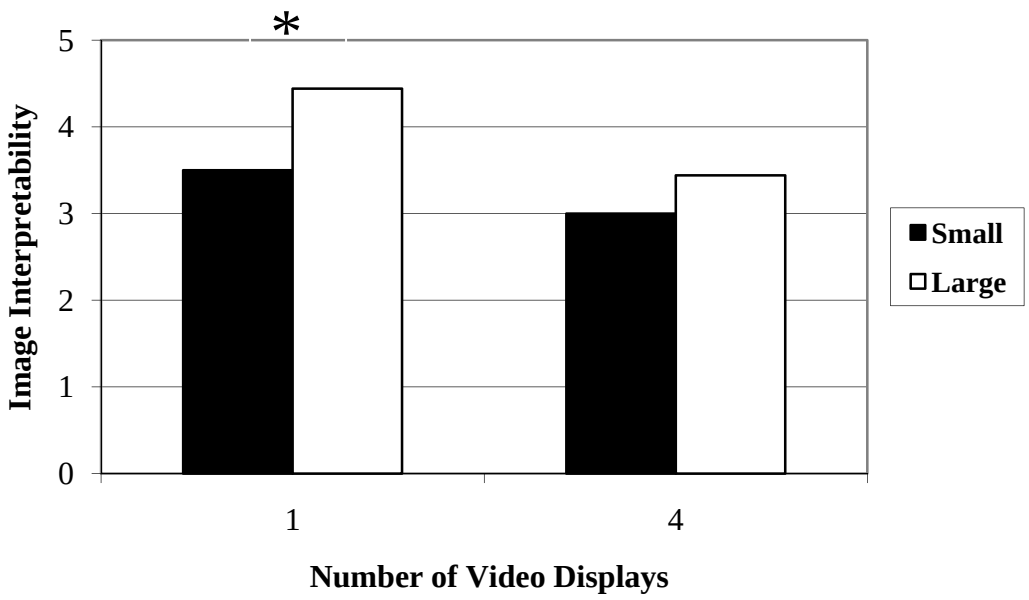


Figure 16. Image interpretability rating by number of videos and video display size.

Target Attributes, Task Characteristics, and Target Detection Accuracy

Post-hoc analyses were performed to examine the relations between target attributes (target size, length of time the target was viewable) and task characteristics (number of video displays) and target detection probability in order to improve our understanding of the factors affecting target detection accuracy. There were 98 targets in all; two single videos with 10 targets each, one four-video condition with 14 targets and one four video condition with 15 targets. Each of these video arrangements was viewed twice, once in the small screen format and once in the large screen format (2 display sizes * (10 + 10 + 14 + 15 targets) = 2 * 49 = 98 targets).

Examination of the descriptive statistics for the target attributes indicted substantial variability for target size and time on screen and for target detection accuracy for the 98 targets. Mean time on screen for all 98 targets was 2.61 seconds, with a minimum of 0.20 seconds and a maximum of 9.92 seconds. Both the minimum and maximum time on screen targets were presented in the one video screen condition. Mean size for the 49 targets in the small display format was 1,393 pixels and ranged from 234 to 9,724 pixels. Mean target size for the 49 targets in the large display format was 5,225 pixels and ranged from 937 to 38,896 pixels. Mean hit rate for all 98 targets was 72.07 percent and ranged from 0.00 percent to 100.00 percent. A closer

examination of the data revealed that one of the 98 targets was never detected and it occurred in the one small video presentation. Also, 13 of the 98 targets were detected 100 percent of the time (3 in the one small video condition, 7 in the one large video condition, 0 in the four small videos condition, and 3 in the four large videos condition).

Both time on screen ($r = 0.525$, $p < 0.01$) and number of video displays ($r = -0.173$, $p < 0.05$) were correlated significantly with target detection accuracy (hit percent), while target size ($r = -0.078$, ns) was not. A regression model that used time on screen, target size, and number of videos was significantly related to target hit percent ($R = 0.538$, $p < .001$). However, this model was not significantly different from one that used only time on screen ($r = 0.525$). These results are consistent with previous empirical studies regarding the probability of target detection in time-limited search (Wilson, Devitt, & Maurer, 2005). Wilson et.al.. demonstrated a strong non-linear mathematical relationship between time available for search and probability to detect a target. In their model, the probability to detect a target was nearly 0 percent when time available to search was less than 0.8 seconds, increased steeply as time approached 3 seconds, then increased at a much lower rate.

Lessons Learned

Results indicated that participants' ability to detect targets was not affected by display size. However, more false alarms occurred for the small display size when four videos were monitored. If display limitations require a small display format, it would be desirable to include a target confirmation step following initial detection. Large display sizes are preferable if target detection without errors is critical and no target confirmation step is included. Follow-on studies should be conducted to examine display design concepts to mitigate false alarms. Target confirmation could involve the operator or a different person review snapshots or short video clips that contain the suspected targets. A potential drawback to adding a confirmation step is significant delays in target reporting time and an increase in misses due to attention required for the confirmation process. Post-hoc analyses revealed that the amount of time an object is viewable is critical to construction of stimulus materials for target detection tasks. Target detection tasks should include targets with varying attributes (i.e., a range of time available for viewing, size).

EXPERIMENT 4: MANUAL VS. COOPERATIVE CONTROL OF WIDE AREA SEARCH MUNITIONS

Method

The United States Air Force is considering advanced automation system concepts that could deploy multiple semi-autonomous unmanned weapons systems into the battle zone. One such system, the Wide Area Search Munitions (WASMs), is a hybrid that combines the attributes of an unmanned aerial vehicle with those of traditional munitions systems. The WASM concept envisions artificially intelligent munitions that communicate and coordinate with one another and human operators to perform their tasks more effectively. Since cooperating WASMs have not yet been produced, research into strategies for controlling them presents a challenging problem that is being approached by simulating WASMs as accurately as possible and evaluating them in human-in-the-loop simulations and concept of operations scenarios (Scerri, Liao, Lai, Sycara, Xu, & Lewis, 2005).

The objective of Experiment 4 was to examine target acquisition performance for unaided human operators with that of an automated cooperative controller in accomplishing a complex task involving the prosecution of ground based targets with WASMs. This purpose was to provide empirical data on an operator's ability to simultaneously manage multiple WASMs while performing a target search, identification, and weapon assignment task. This information will provide valuable insights into concepts of employment and technology requirements for future munitions and semi-autonomous systems (e.g., how much automation is acceptable, information requirements, need for decision aiding software, manpower and personnel qualification requirements). See Carretta, Warfield, and Patzek (2009) for a summary.

Participants

Twelve full-time civilian and military employees stationed at Wright-Patterson AFB OH participated in this study. This sample consisted of 12 men who ranged in age from 20 to 45 years with a mean of 30.3 years. All participants reported being in good to excellent health and having vision correctable to 20/20, normal color vision, and normal peripheral vision. Most participants indicated that they had previous simulator (67 percent) and video game (92 percent)

experience. Participation was voluntary and no compensation was offered in exchange for participation in the experiment.

Materials

Laser Detection and Ranging (LADAR) Imagery

The WASM platforms transmit LADAR imagery to operators who must interpret it to make target acquisition decisions. LADAR is similar to millimeter wave radar, but employs laser beams to scan and process the signal echoed from targets to create a 3-D virtual picture of the area. Simulated LADAR imagery was used in the current experiment. Participants viewed still LADAR images of each designated target showing a top, side, front, and back view of the targets.

Data Collection Instruments

Task performance and questionnaire data were collected. The questionnaires are described below and are available in Warfield, Carretta, Patzek, O'Neal, and Estepp (in press).

Objective measures of target acquisition performance. Several objective measures of target acquisition performance were collected. These were number of high priority targets attacked, number of low priority targets attacked, mean time on target, mean time on target error, standard deviation of time on target, time to plan, and time to complete (Table 9).

Demographic data questionnaire. This questionnaire (Appendix A) is the same as that used in Experiments 1 and 3.

Confidence ratings. At the completion of each target acquisition/weapon assignment scenario, participants were instructed to indicate the level of confidence in their target acquisition decisions. Confidence ratings (Appendix I) were made on a five-point Likert rating scale (1 - not at all confident, 2 - slightly confident, 3 - moderately confident, 4 - fairly confident, 5 – very confident).

NASA-TLX. This measure (Appendix F) is the same as that described in Experiments 2 and 3.

Table 9. Objective Measures of Task Performance

Measure	Definition
Number of High Priority Targets Attacked	Mean number of high priority targets attacked
Number of Low Priority Targets Attacked	Mean number of low priority targets attacked
Mean Time on Target	The average time on target for the WASMs
Mean Time on Target Error	The average error between the time on target and requested time on target. That is, how close the attacks were to the requested time. This score could be computed only for the cooperative control condition.
Standard Deviation of Time on Target	This is the standard deviation of the actual time on target compared with mean time on target (i.e., how close the attacks were to each other).
Time to Plan	Time from when the first target was selected to attack authorization or cancellation.
Time to Complete	Time from authorization to when the last target is attacked.

Post-test questionnaire. This questionnaire elicited information regarding participants' assessment of the operator interface. Participants rated the operator interface for ease of use to identify the targets and classify their priority level (high or low). Participants also provided a self-assessment of their ability to perform a near simultaneous attack under the manual and cooperative control conditions. These questions used a five-point scale (1 - poor, 2 - fair, 3 - good, 4 - very good, 5 – excellent). Participants also were given the opportunity to provide comments regarding the operator interface and other factors that affected their ability to identify, classify, and attack targets.

Equipment

Figure 17 shows a test participant interacting with the WASM test station. Experiment 4 was conducted in the Crew Station Integration Laboratory in the 711th Human Performance Wing, Supervisory Control Interfaces Branch. Participants were seated in a non adjustable crew member's chair attached to rails. The chair was located in the aft end of a generic cargo aircraft simulator. Participants were seated directly in front of a 13.3 inch CF-73 Panasonic laptop that presented the simulated WASMs attacking targets on a Falcon View map. Still images of potential targets were displayed on a poster next to the laptop computer to aid the participants during target acquisition. Participants used a mouse with a scroll wheel to designate target detection and weapon assignment decisions. A laptop computer was placed nearby where participants entered questionnaire responses.

Figure 17. WASM experimental station.

Procedures

The experimental session began with a pre-briefing, participant informed consent, and completion of a short biographical questionnaire. The pre-briefing provided information regarding the purpose of the study, equipment, controls, and displays to be used, procedures, and the mission scenario. Participants remained in a seated position during the pre-briefing, practice, and data collection.

Following the pre-briefing, training was conducted to achieve familiarity with test equipment, procedures, and tasks. Participants completed three practice trials for each level of control (manual vs. cooperative control) by number of WASMs (4, 8, or 16) combination using a representative target set.

Prior to starting the test trials, participants were fitted with physiological electrodes to measure electrical brain, eye, and heart activity.[1] There were nine test trials for each level of control by number of WASMs combination. Immediately following each test trial, participants rated the level of confidence in their target acquisition decisions and subjective workload. After completion of the final test session, participants completed the post-test questionnaire regarding their experience.

Analyses

The purpose of the study was to compare the objective and subjective data on a target acquisition task for manual versus cooperative control over three levels of mission complexity (4, 8, or 16 WASMs). Related samples t-tests and repeated measures analyses of variance were performed since participants were exposed to all level of control by number of WASMs combinations. Partial eta squared and observed power were reported in conjunction with the analyses of variance. Partial eta squared is a measure of effect size. It is the proportion of the effect plus error variance that is attributable to the effect; thus, the larger the value the more variance that is explained by the effect (e.g., level of control, number of WASMs). The power of a statistical test is the probability that the test will reject a false null hypothesis (will not make a Type II error). As power increases, the probability of a Type II error decreases. Observed power is computed after the study has been completed and uses the obtained sample size and effect size to determine what the power was in the study, assuming the effect size in the study is equal to that in the population. As with partial eta squared, the larger the value the better.

Objective measures of performance included hit accuracy (percent), number of false alarms, and amount of time required to attack all targets. Subjective measures were participants'

[1] The physiological data had not been processed and analyzed in time for inclusion in this report.

overall workload, confidence in target acquisition decisions and their self-assessment of the ability to accomplish near simultaneous attack.

It was assumed that task difficulty would increase going from cooperative control to manual control and as the number of WASMs increased from 4 to 8 to 16. As a result, all analyses were performed using a 0.05 Type I error rate and a directional hypotheses.

Experiment 4: Results and Discussion

Target Acquisition

Number of Hits

It was expected that performance under the cooperative control mode would equal or exceed that under the manual control mode, so one-directional hypotheses were tested. Comparisons between the cooperative control and manual control modes indicated that within each number of WASMs condition, there was no significant decrement in the number of high priority targets attacked. However, for the number of low priority targets attacked more targets were attacked for the cooperative control mode in the 16 WASMs condition (3.69 vs. 3.41; $t = 2.41$, $p \leq 0.05$)(Table 10).

Time on Target, Time to Plan, and Time to Complete Measures (Table 11)

For *Mean Time on Target* (i.e., the average of the actual time on target for the WASMs), no significant effects were observed for level of control, number of WASMs, or their interaction.

Mean Time on Target Error (i.e., how close the attacks were to the requested time) generally increased as the number of WASMs/targets increased (4 WASMs = 2.04, 8 WASMs = 1.30, 16 WASMs = 8.58). The low value for the 8 WASM condition may have occurred due to the closer placement of targets in this condition relative to the 4 WASM/targets condition. It should be noted that mean time on target error cannot be computed for the manual mode because a requested time on target cannot be specified in manual mode.

SD Time on Target Error (i.e., how close the attacks were to each other) was significantly affected by level of control, number of WASMs/targets, and their interaction. An examination of

the means showed that time between attacks was greater for the manual versus cooperative control condition and generally increased as the number of WASMs/targets increased.

Table 10. Number of Hits: Cooperative Control versus Manual Mode

		Cooperative Control		Manual			
Score	N WASMs	Mean	SD	Mean	SD	df	t
N High Priority Hits	4	3.33	0.00	3.27	0.12	11	1.48
	8	6.66	0.00	6.55	0.38	11	1.00
	16	12.30	0.09	12.52	0.33	11	-2.00
N Low Priority Hits	4	0.66	0.00	0.69	0.17	11	-0.56
	8	1.33	0.00	1.33	0.14	11	0.00
	16	3.69	0.09	3.41	0.35	11	2.41*

N = 12; *p≤ .05

Significant effects were observed for both *Time to Plan* and *Time to Complete* for control mode and number of WASMs/targets. *Time to Plan* was greater for manual control ($F(1, 11) = 20.70$, $p < .01$) and increased as the number of WASMs/targets increased ($F(2, 10) = 19.76$, $p < .01$). *Time to Complete* was *less* for manual control ($F(1, 11) = 490.81$, $p < .01$) and increased as the number of WASMs/targets increased ($F(2, 10) = 6.89$, $p < .01$). At first, it appears counterintuitive that *Time to Complete* was lower for the manual versus the cooperative control mode. However, it should be noted that in the manual control mode, target authorization and attack occur separately for each WASM/target combination and once authorization has occurred, the WASM takes a direct flight path to the target. In the cooperative control mode the attack does not occur until all target/WASM combinations have been authorized and it is necessary for some WASMs to employ longer flight paths to enable simultaneous attack.

\

Table 11. Means and Standard Deviations: Time on Target, Time to Plan, and Time to Complete Scores

		Cooperative Control		Manual	
Score	N WASMs	Mean	SD	Mean	SD
Mean Time on Target	4	494.00	83.88	573.84	327.90
	8	488.57	55.83	446.71	67.35
	16	540.15	75.55	552.56	288.37
Mean Time on Target Error	4	2.04	1.22	-------	-------
	8	1.30	0.53	-------	-------
	16	8.58	4.44	-------	-------
SD Time on Target Error	4	2.24	2.11	10.17	4.21
	8	1.45	1.44	17.58	7.16
	16	9.09	6.16	27.43	11.89
Time to Plan	4	22.47	4.00	39.40	15.66
	8	36.01	7.63	61.26	26.83
	16	70.16	13.71	105.24	51.05
Time to Complete	4	117.22	11.89	63.06	10.45
	8	124.63	7.49	65.64	5.43
	16	148.09	26.76	74.96	10.90

N = 12

Confidence Ratings

Although there was a trend toward greater confidence for decisions made using the cooperative control mode, this trend was not statistically significant. It should be noted that the observed power for this test was low, suggesting that if a larger sample were tested the effect

might reach statistical significance. Mean confidence level was related significantly to the number of WASMs/targets. An examination of the means showed a general trend toward lower confidence as the number of WASMs increased, especially for the manual control mode.

Workload

Subjective workload was measured using the NASA TLX. As previously discussed, the NASA TLX has 6 subscales that are combined to create an overall workload index. Examination of the means revealed a consistent trend toward increased workload going from the cooperative control mode to manual control mode and from 4 to 8 to 16 WASMs. This trend was statistically significant for the Total workload score and for all of the NASA TLX scales except Physical workload.

Post-Test Questionnaire

Following completion of the test trials, participants completed a post-study questionnaire regarding their experience. They rated ease with which they were able to use the operator interface to identify targets and their ability to classify the priority level of targets using the WASM interface. Both ratings were on a 5 point scale: 1 – poor, 2 – fair, 3 – good, 4 – very good, and 5 – excellent. Although ratings for ease of use and ability to classify the target priority level varied, the mean ratings for both approached “very good.” Rating for ease of use ranged from 3 to 5 with a mean of 3.92; those for ability to classify the target priority level ranged from 2 to 5 with a mean of 3.83.

Participants then rated their ability to perform a simultaneous attack using the cooperative control and manual control modes for the 4 and 16 WASM/target conditions (Table 12). Ratings were on a five point scale: 1 – poor, 2 – fair, 3 – good, 4 – very good, and 5 – excellent. Inspection of the means showed a strong trend toward lower ratings of ability to perform a simultaneous attack for the manual control mode and for the 16 WASMs/targets condition. The effect was especially pronounced for the manual control mode condition with 16 WASMs/targets (mean = 1.5).

Table 12. Means and Standard Deviations: Ability to Perform Simultaneous Attack

		Cooperative Control		Manual	
Score	N WASMs	Mean	SD	Mean	SD
Ability to Perform Simultaneous Attack	4	4.83	0.389	4.17	0.178
	16	3.83	0.835	1.50	0.674

N = 12

Participants had the opportunity to provide open-ended comments regarding the WASM interface and procedures. Seven of the 12 participants made one or more comments. These focused on ways to improve the manual control mode and the interface design. Suggestions regarding the manual control mode included adding the ability to insert waypoints and timing points to improve simultaneous attack. Suggestions regarding the interface design focused on providing multiple data input options in addition to the mouse and using a larger screen or multiple screens.

Lessons Learned

Participants were able to acquire and attack nearly all of the targets even under the most demanding condition, that is, manual control of 16 WASMs. As expected, unaided operators were not able to achieve simultaneous attack of the targets as efficiently as the cooperative controller. Time between attacks was greater for the manual versus cooperative control mode and generally increased as the number of WASMs/targets increased. The decrement in performance efficiency between the manual and cooperative control modes is important under the circumstance when it is crucial to limit the amount of time an adversary has to respond to a first attack. Even in the least demanding condition involving 4 WASMs/targets, participants' ability

to manually perform a near simultaneous attack was degraded compared to the cooperative control mode. These results are also reflected in participants' self-assessments of workload and their ability to perform a near simultaneous attack.

Additional studies are needed to examine factors that may affect performance differences between the manual and cooperative control modes. For example, the extent to which targets are clustered (or dispersed) in the search area may affect the relative efficiency of the manual and cooperative control modes. Also, it would be informative to examine additional numbers of WASMs/targets (1, 2, 3, … n) to better determine performance differences between the manual and cooperative control modes.

FUTURE DIRECTIONS

Although the four ISUS experiments provided valuable insights into factors affecting the ability of human operators to perform a target acquisition task while monitoring one or more unmanned air vehicles, the results raised as many questions as they answered. For example, there are many issues regarding the effectiveness of alternate presentation methods (e.g., video frame rate, live versus mosaic-based or DVR presentation) and decision aids (e.g., automatic target cueing, anomaly detection algorithms, interruption recovery methods) for enhancing target acquisition performance. Future studies will build on prior research and will develop, expand, and optimize situation assessment and decision support information displays for multi-UAV control. This is to be achieved through user-centered approaches to system design and assessment, including cognitive engineering methods, usability assessments and evaluation of operator models. Using information feeds expected to be available for future operations (blue force tracking, improved weather feeds, etc), this effort will research novel information fusion and management concepts (and associated metrics) that maximize effectiveness through intuitive information transfer. The goals of this line of research are to reduce the operator to vehicle ration (i.e., increase span of control, reduce deployment footprint), increase mission effectiveness (i.e., improve time critical operations, increase mission flexibility), and improve awareness of system state/intent (i.e., timely response to contingencies, reduced mishaps).

REFERENCES

Alexander, A. L., Nygren, T. E., & Vidulich, M. A. (2000). *Examining the relationship between mental workload and situation awareness in a simulated air combat task,* AFRL-HE-WP-TR-2000-0094. Air Force Research Laboratory, Human Effectiveness Directorate, Wright-Patterson AFB, OH.

Barbato, G. (2000). Uninhabited combat air vehicle controls and displays for suppression of enemy air defenses. *CSERIAC Gateway. 11 (1),* 1-4.

Carretta, T. R., Warfield, L., & Patzek, M. J. (2009). Manual and cooperative control mission management methods for wide area search munitions. *Proceedings of the 15th International Symposium on Aviation Psychology.* Dayton, OH.

Clough, B. T. (2002). *UAV swarming? So what are those swarms, what are the implications, and how do we handle them?,* AFRL-VA-WP-TP-2002-308. Wright-Patterson AFB, OH: Air Force Research Laboratory, Air Vehicles Directorate, Control Sciences Division.

Corriveau, P., Gojmeiac, C., Hughes, B., & Stelmach, L. (1999). All subjective scales are not created equal: The effects of context on different scales. *Signal Processing, 77,* 1-9.

Dixon, S. R., & Wickens, C. D. (2003). Control of multiple UAVs: A workload analysis. *Proceedings of the 12th International Symposium on Aviation Psychology,* Dayton, OH.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA TLX (task load index): Results of empirical and theoretical research in human mental workload. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139-183). New York: Elsevier.

Irvine, J. M. (July, 1997). National Imagery Interpretability Rating Scales (NIIRS): Overview and methodology. *SPIE, Airborne Reconnaissance XXI, 3128,* 93-103.

Irvine, J. M., Fenimore, C., Cannon, D., Roberts, J., Israel, S. A., Simon, L., Watts, C., Miller, J. D., Brennan, B., Aviles, A. I., Tighe, P. F., & Behrens, R. J. (2005). Feasibility study for the development of a motion image quality metric. *Proceedings of the 33rd Applied Imagery and Pattern Recognition Workshop: Image and Data Fusion, IEEE Computer Society* (pp. 179-183), Washington, DC.

Irvine, J. M., Fenimore, C., Cannon, D., Roberts, J., Israel, S. A., Simon, L., Watts, C., Miller, J. D., Brennan, B., Aviles, A. I., Tighe, P. F., Behrens, R. J., & Haverkemp, D. (2005). Factors affecting development of a motion imagery quality metric. *Proceedings of SPIE Visual Information Processing XIV, volume 5817,* (pp. 116-123). Orlando, FL.

Janson, W. P., See, J. E., Riegler, J. T., Davis, I., & Kuperman, G. G. (1998). *Aided and unaided operator performance with first generation FLIR imagery,* AFRL/HE-WP-TR-1999-0003. Wright-Patterson AFB, OH: Human Effectiveness Directorate, Human Interface Division.

Leachtenauer, J. (1996). National Imagery Interpretability Rating Scale: Overview and product description. *Proceedings of the annual meeting of the American Society of Photogrammetry and Remote Sensing.*

Maver, L. M., Erdman, C. D., & Riehl, K. (1995). *Imagery interpretability rating scales.* Paper presented at the meeting of the Society for Information Display.

Office of the Secretary of Defense (2001). *Unmanned aerial vehicles roadmap: 2000-2025*. Washington, DC.

Petkosek, M. A., Carretta, T. R., Patzek, M. J., & Stoor, B. J. (2007). *Multi-aircraft video-human/automation target recognition: Aided/unaided target acquisition involving multiple simulated micro air vehicles,* AFRL-HE-WP-TR-2007-0098. Wright-Patterson AFB, OH: Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter Interface Division.

Plantz, S. E., Warfield, L., Carretta, T. R., Gonzalez-Garcia, A., & Patzek, M. J. (2008). *Multi-aircraft video- human/automation target recognition studies: Video display size in unaided target acquisition involving multiple videos,* AFRL-RH-WP-TR-2008-0074. Wright-Patterson AFB, OH: Air Force Research Laboratory, Human Effectiveness Directorate, Warfighter Interface Division.

Riehl, K., & Maver, L. (1996). A comparison of two common aerial reconnaissance image quality measures. *SPIE, Airborne Reconnaissance XX, 2829,* 242-254.

Ruff, H. A., Calhoun, G. L., Draper, M. H., Fontejon, J. V., Guilfoos, B. J. (2004). Exploring automation issues in supervisory control of multiple UAVs. *Proceedings of the Human Performance Situation Awareness, and Automation Technology Conference,* 218-222.

Scerri, P., Liao, E., Lai, J., Sycara, K., Xu, Y., & Lewis, M. (2005). Coordination very

large groups of wide area search munitions. In D. Grundel, R. Murphey, & P. M. Partdalos (Eds.), *Series on Computers and Operations Research, vol. 4, Theory and Algorithms for Cooperative Systems* (pp. 1-31). World Science Publishing Company.

See, J. E., Davis, I., & Kupperman, G. G. (1997). *Automatic target cueing and operator performance with enhanced APG-70 synthetic aperture radar imagery,* AL/CF-TR-1997-0171. Wright-Patterson AFB, OH: Armstrong Laboratory.

USA Today (January 1, 2008). *Military use of unmanned aircraft soars.* http://www.usatoday.com/news/military/2008-01-01-unmanned-killers_N.htm

Walker, B. (2005). *Virtual environment UAV swarm management using GPU calculated digital pheromones.* Unpublished doctoral dissertation, Iowa State University, Ames, IO.

Warfield, L., Carretta, T. R., Patzek, M. J., & Gonzalez-Garcia, A. (2007). *Multi-aircraft video-human/automation target recognition studies: Unaided target acquisition involving multiple micro air vehicle (MAV) videos,* AFRL-HE-WP-TR-2007-0036. Wright-Patterson AFB, OH: Air Force Research Laboratory, Warfighter Interface Division.

Warfield, L., Carretta, T. R., Patzek, M. P., O'Neal, J. K. & Estepp. J. (in press). *Comparing manual and cooperative control mission management methods for wide area search munitions,* AFRL-RH-WP-TR-2009-xxxx. Wright-Patterson AFB, OH: Air Force Research Laboratory, Warfighter Interface Division.

Wickens, C. D., Dixon, S.R., & Chang, D. (2003). *Using interference models to predict performance in a multiple-task UAV environment – 2 UAVs,* Technical Report AHFD-03-9/MAAD-03-1. Micro Analysis and Design, Boulder, CO.

Wilson, D., Devitt, N., & Maurer, T. (2005). Search times and probability of detection in time-limited search. Infrared imaging Systems: Design, Analysis, Modeling, and Testing XVI, G. C. Holst (Ed.). *Proceedings of SPIE , 5784,* 224-231.

Appendix A: Demographic Questionnaire for Experiment #1

Demographic Data Questionnaire – Experiment #1

Participant ID: _____

1. Age: _____

2. Gender (circle one) Male Female

3. Describe your general health (circle one):

   Poor Fair Good Very Good Excellent

4. How would you assess your overall feeling of wellbeing this morning/afternoon (circle one)?

   Poor Fair Good Very Good Excellent

5. Do you have any practical experience working in a simulation type environment?

   If yes explain:

6. Do you play any type of computer/video games? Yes No
   a. If you answered "Yes," what types do you play? (circle all that apply)

   Action/Adventure _____ Role Playing _____ Other (specify) _____

   b. Do the computer/video games you play require you to do visual search tasks (i.e., locate/identify objects or targets)? Yes No

7. Is your visual acuity correctable to 20/20? Yes No

8. Do you have any problems with your peripheral vision? Yes No

9. Are you color blind? Yes No

10. Are you aware you may withdraw from this study at any time? Yes No

11. Are you aware that your participation is strictly confidential? Yes No

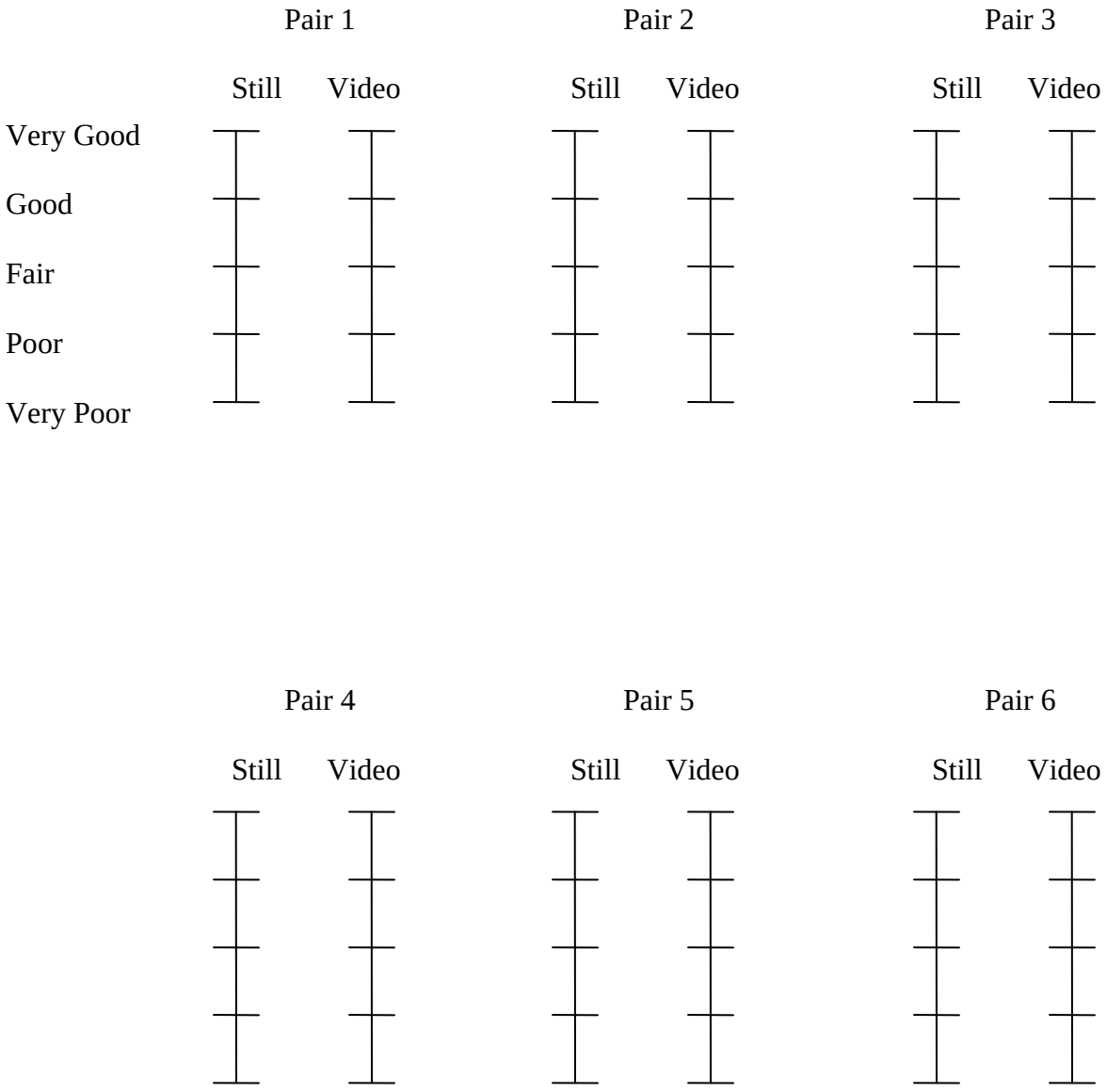
Appendix B : Civil National Imagery Interpretability Rating Scale

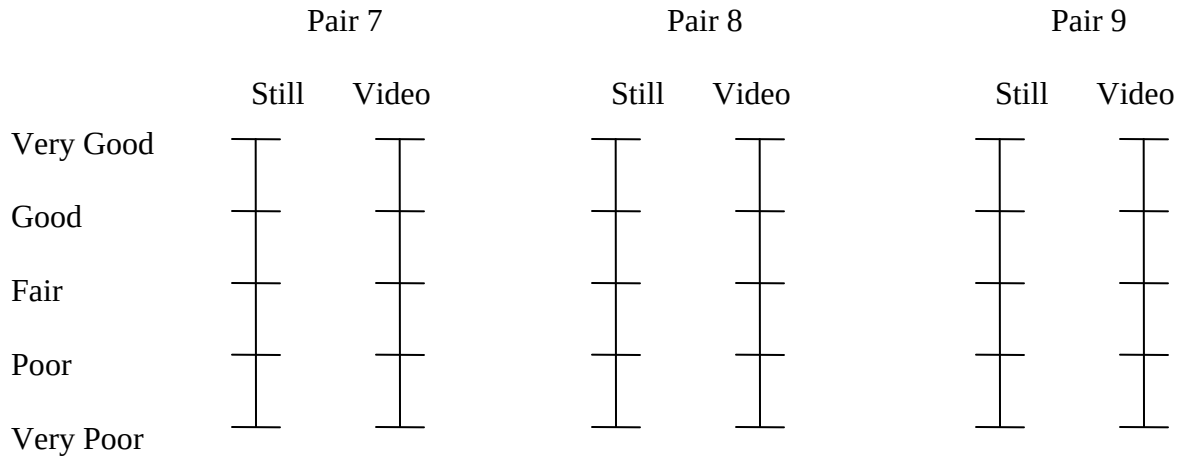
Civil National Imagery Interpretability Rating Scale

Rating Level	Descriptive Examples
0	Interpretability of imagery is precluded by obscuration, degradation, or very poor resolution.
1	Distinguish between major land use classes (e.g., agricultural, barren, forest, urban, water). Detect a medium-sized port facility. Distinguish between runways and taxiways at a large airfield. Identify large area drainage patterns by type (e.g., dendritic, radial, trellis).
2	Identify large (i.e., greater than 160 acres) center-pivot irrigated fields during the growing season. Detect large buildings (e.g., factories, hospitals). Identify road patterns like cloverleafs, on major highway systems. Detect ice-breaker tracks. Detect the wake from a large (e.g., greater than 330 ft.) ship.
3	Detect large area (e.g., greater than 160 acres) contour plowing. Detect individual houses in residential neighborhoods. Detect trains or strings of standard rolling stock on railroad tracks (not individual cars). Identify inland waterways navigable by barges. Distinguish between natural forest stands and orchards.
4	Identify farm buildings as barns, silos, or residences. Count unoccupied railroad tracks along right-of-way or in a railroad yard. Detect basketball court, tennis court, or volleyball court in urban areas. Identify individual tracks, rail pairs, control towers, switching points in a rail yard. Detect jeep trails through grassland.
5	Identify Christmas tree plantations. Detect open bay doors of vehicle storage buildings. Identify tents (larger than 2-person) at established recreational camping areas.

	Distinguish between stands of coniferous and deciduous trees during leaf-off condition. Detect large animals (e.g., elephants, giraffes, rhinoceros) in grasslands.
6	Detect narcotics intercropping based on texture. Distinguish between row crops (e.g., corn, soybean) and small grain crops (e.g., barley, oats, wheat). Identify automobiles as sedans or station wagons. Identify individual telephone/electric poles in residential neighborhoods. Detect foot trail through barren neighborhoods.
7	Identify individual mature cotton plants in a known cotton field. Identify individual railroad ties. Detect individual steps on a stairway. Detect stumps and rocks in forest clearings and meadows.
8	Count individual baby pigs. Identify a USGS benchmark set in a paved surface. Identify grill detailing and/or the license plates on a passenger/truck type vehicle. Identify individual pine seedlings. Identify individual water lilies on a pond, Identify windshield wipers on a vehicle.
9	Identify individual grain heads on small grain (e.g., barley, oats, wheat). Identify individual barbs on a barbed wire fence. Detect individual spikes in railroad ties. Identify individual bunches of pine needles. Identify an ear tag on large animals (e.g., deer, elk, moose).

Appendix C : Image Quality Rating Scale


Pair 1
Pair 2
Pair 3
Still
Video
Still
Video
Still
Video
Very Good
Good
Fair
Poor
Very Poor
Pair 4
Pair 5
Pair 6
Still
Video
Still
Video
Still
Video


Pair 7
Pair 8
Pair 9
Still
Video
Still
Video
Still
Video
Very Good
Good
Fair
Poor
Very Poor

Appendix D: Post-Test Interview Questionnaire for Experiment #1

Post-Test Interview Questions – Experiment #1

1. How would you rate the quality of the video presented on this display device? (circle one)

Poor Fair Good Very Good Excellent

If you answer to #1 was “poor” or “fair,” what factors affected your rating?

2. How would you assess the clarity of the video imagery? (circle one)

Poor Fair Good Very Good Excellent

3. How would you assess the contrast of the video imagery? (circle one)

Poor Fair Good Very Good Excellent

4. How would you assess the resolution of the video imagery? (circle one)

Poor Fair Good Very Good Excellent

5. Did the display provide a sufficiently interpretable image? (circle one) Yes No

6. Were you able to identify all predefined targets of interest in the video? (circle one)

Yes No

If no explain:

7. Please provide any additional comments below:

Appendix E: Demographic Questionnaire for Experiment #2

PARTICIPANT DEMOGRAPHICS – EXPERIMENT #2

1. Participant number:_____

2. Today's date:_____

3. Your birth date (MM/DD/YYYY):_____

4. Preferred Handedness: (please circle one)
   Left Right

5. Do you feel comfortable using a computer mouse: (please circle one)
   Yes No (if yes, please answer 5b; if no, please go to 6)

   5b. Which hand do you consistently use the mouse with: (please circle one)
   Left Right

6. Duty station (office symbol) and Job series number:_____ / (job)_____

7. Please describe your job duties:
_____
_____
_____
_____
_____

8. Is your vision correctable to 20/20? (please circle one)
   without contacts/glasses with contacts/glasses

9. Do you have normal color vision: (please circle one)
   Yes No

10. By estimation, what is the average amount of time you use a computer before taking a break?
:_____

11. Do you use a computer for any routine activities: (please circle one)
   Yes No (if yes, please answer 11b and c; if no, please go to 12)

   11b. Where do you use a computer: (please check <u>all</u> that apply)
   ___ Home
   ___ Work
   ___ Other: _____

   11c. For each of these locations, please indicate the types of applications you use:

Home:

_____
_____

Work:

_____
_____

Other:

_____
_____

12. Have you played video/computer games regularly since January 2006 until the present: (please circle one)
Yes No (if yes, please answer 12b; if no, please go to 13)

12b. What kinds of games have you played:
(please check all that apply)
___ Role playing games (RPGs: Final Fantasy, Myst, EverQuest, etc)
___ Action (SOCOM Navy Seals, Rainbow 6, Counter Strike, etc)
___ Sports (EA Sports NCAA College Football, NBA Jam, EA Sports MVP Baseball, etc)
___ Special Interest (Casino style games, puzzles, etc)
___ Other: _____

C2

13. Are you familiar with Air Force Concept of Operations for small and micro UAVs: (please circle one)
Yes No (if yes, please explain your experiences and knowledge of these systems that is not restricted by classification, FOUO, sensitivity, or proprietary rights)

_____
_____

If applicable, "I cannot discuss my knowledge and experience with small and micro UAVs," please check this box □

Appendix F: NASA-TLX Instructions and Questionnaire

NASA-TLX

We are not only interested in assessing your performance but also the experiences you have during the experimental trials. Right now we are going to describe the technique that will be used to examine these experiences. In the most general sense we are examining the “Workload" you experience. Workload is a difficult concept to define *precisely*, but a simple one to understand *generally*. The factors that influence your experience of workload may come from the task itself, your feelings about your own performance, how much effort you put in, or the stress and frustration you feel. Physical components of workload are relatively easy to conceptualize and evaluate. However, mental components of workload may be more difficult to measure.

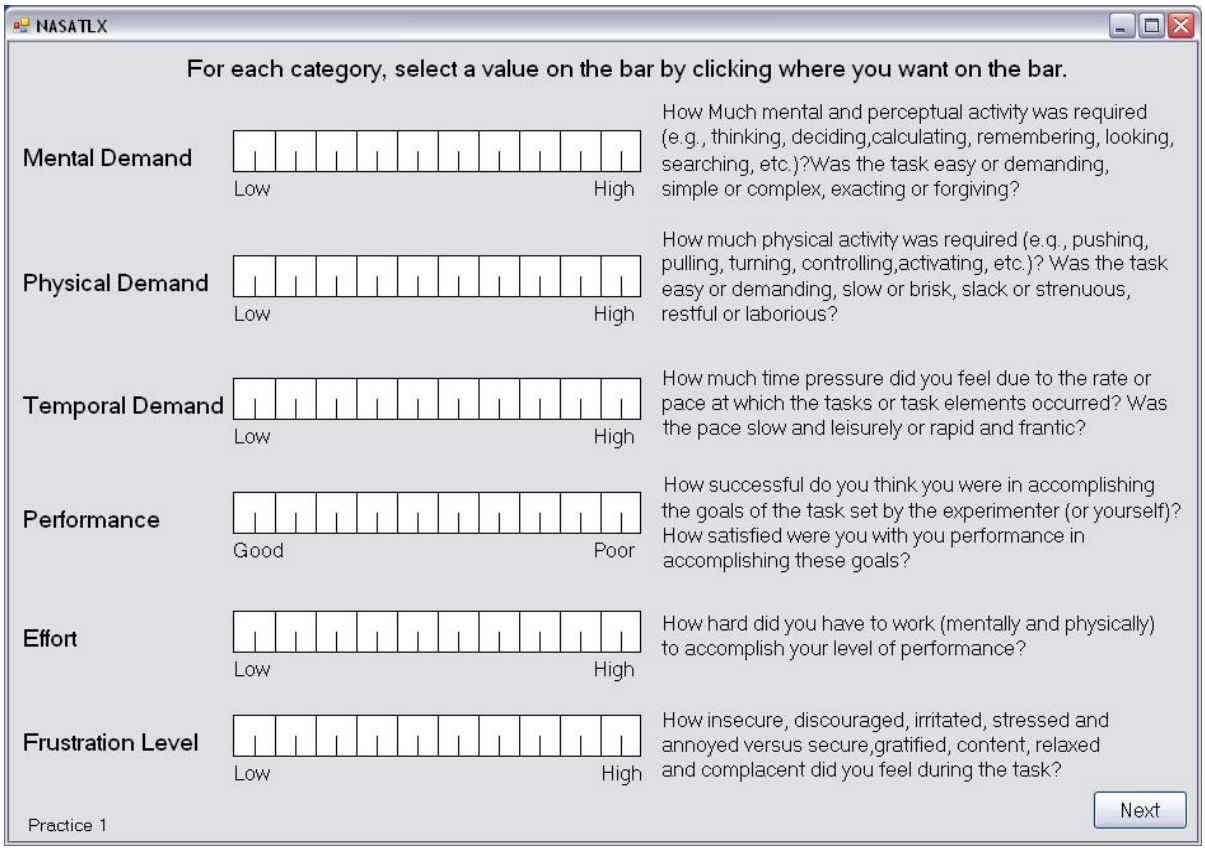
Since workload is something that is experienced individually by each person, there are no effective "rulers" that can be used to estimate the workload of different activities. One way to find out about workload is to ask people to describe the feelings they experienced. Because workload may be caused by different factors, we would like you to evaluate several of them individually rather than lumping them into a single, global evaluation of overall workload. This set of six rating scales was developed for you to use in evaluating your experiences during different tasks. (Hand scale sheet on top of explanations to participant)

Please read the descriptions of the scales carefully. If you have a question about any of the scales in the table, please ask me about it. It is important that they be clear to you. You may keep the descriptions with you for reference during the experiment.

(Stop here, read detailed subscale explanations while participant reviews the scale sheet/explanations)

After performing each task, you will evaluate it by marking each scale at the point that matches your experience. Each line has two endpoint descriptors that describe the scale. Note that "performance" goes from “good" on the left to “poor" on the right. This order has been confusing for some people. Mark the desired location. Please consider your responses carefully in distinguishing among the task conditions. When rating each task, only reflect on the one you have just completed. Consider each trial in isolation, that is, do not compare it to prior experiences. Also, please consider each scale individually. Although the definitions may be similar for two or more scales, try to distinguish them from each other based on my explanations and the definitions that you may refer to throughout the experiment- even when rating them.

Your ratings will play an important role in the evaluation being conducted, thus, your active participation is essential to the success of this experiment, and is greatly appreciated!



NASA-Task Load Index computer application used. Participants rated their perception of the workload for each individual trial in the six categories, descriptions of the categories are located to the right of the scale (top image). When finished with ratings the participant completed a pair wise comparison of the categories (bottom image).

Appendix G: Post-Test Interview Questionnaire for Experiment #2

POST-EXPRIMENT SUBJECTIVE QUESTIONNAIRE – EXPERIMENT #2

1. How difficult was the target detection task?
(please circle one number)

Less Difficult 1 2 3 4 5 More Difficult

2. What proportion of the targets do you think you correctly identified (please indicate a percentage between 0percent and 100percent) _____

3. Was an automatic target cueing capability used during the study? (please circle one)

Yes No

If, "Yes," please complete 3a – 3d:
If, "No," please complete 4a - 4c:

Before proceeding, please read these definitions. Their principles will be applied in the following questions.

Hit- A vehicle is a target and it is indicated by the automatic target cuer.
Miss- A vehicle is a target and it *is not* indicated by the automatic target cuer.
False Alarm- A vehicle is not a target but it is indicated as a target by the automatic target cuer.

Thank you. Please proceed to the appropriate section of questions, either 3a-3d or 4a-4c.

3a. How reliant were you upon the target cueing? (please circle one number)

Less Reliant on ATC 1 2 3 4 5 Less Reliant on ATC

3b. What did you find drew your attention more?
(please circle one)

Misses False Alarms Equally noticeable

3c. What did you find harder to deal with?
(please circle one)

Misses False Alarms Equally impacted performance

3d. Would you feel confident in using the automatic target cuing capability presented today in real world operations similar to this study? Please disregard "look and feel," consider only effectiveness. (please circle one)

Yes No

4a. How much confidence do you have in your target identification ability? (please circle one number)

Less ← 1 2 3 4 5 → More
Confident in Myself Confident in Myself

4b. How accurate would an automatic target cuer need to be for you to rely on it to make target acquisition decisions?
(please indicate a whole percentage of targets correctly identified between 0percent and 100percent) _____

4c. What percentage of false alarms would be tolerable for an automatic target cuer before you would not trust its recommendations? (please indicate a whole percentage of false alarms between 0percent and 100percent)
_____

You have completed this questionnaire. Thank you for your input!

Appendix H: Post-Test Interview Questionnaire for Experiment #3

Post-Test Interview Questions – Experiment #3

1. How would you rate the quality of the video presented on this display device? (circle one)

Poor Fair Good Very Good Excellent

If your answer to #1 was “poor” or “fair,” what factors affected your rating?

2. How would you assess the clarity of the video imagery? (circle one)

Poor Fair Good Very Good Excellent

3. How would you assess the contrast of the video imagery? (circle one)

Poor Fair Good Very Good Excellent

4. How would you assess the resolution of the video imagery? (circle one)

Poor Fair Good Very Good Excellent

5. Did the display provide a sufficiently interpretable image? (circle one) Yes No

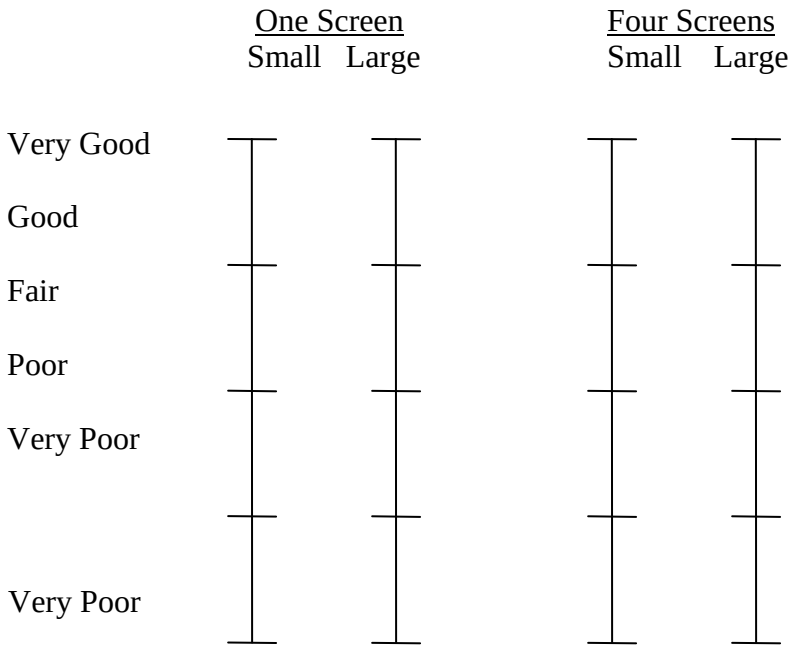
If your answer to #5 was “No,” what factors affected your rating?

6. Were you able to identify all predefined targets of interest in the video? (circle one)

Yes No

If no explain:

7. Please rate the video interpretability of the single screen small and large video displays and the four screen small and large video displays:


One Screen
Small Large
Four Screens
Small Large
Very Good
Good
Fair
Poor
Very Poor
Very Poor

Appendix I: Wide-Area Search Munitions Confidence Questionnaire – Experiment #4

Participant:_____
Date: _____

Four WASM Condition

1.) Run: _____

How confident are you that you attacked the appropriate high and low priority targets?
Please rank your decision between 1 and 5 with (1- not at all confident, 2- slightly confident, 3- moderately confident, 4 - fairly confident, 5 – very confident).

2.) Run: _____

How confident are you that you attacked the appropriate high and low priority targets?
Please rank your decision between 1 and 5 with (1- not at all confident, 2- slightly confident, 3- moderately confident, 4 - fairly confident, 5 – very confident).

3.) Run: _____

How confident are you that you attacked the appropriate high and low priority targets?
Please rank your decision between 1 and 5 with (1- not at all confident, 2- slightly confident, 3- moderately confident, 4 - fairly confident, 5 – very confident).

4.) Run: _____

How confident are you that you attacked the appropriate high and low priority targets?
Please rank your decision between 1 and 5 with (1- not at all confident, 2- slightly confident, 3- moderately confident, 4 - fairly confident, 5 – very confident).

Eight WASM Condition

Participant:_____
Date: _____

1.) Run: _____

How confident are you that you attacked the appropriate high and low priority targets?
Please rank your decision between 1 and 5 with (1- not at all confident, 2- slightly confident, 3- moderately confident, 4 - fairly confident, 5 – very confident).

2.) Run: _____

How confident are you that you attacked the appropriate high and low priority targets?

Please rank your decision between 1 and 5 with (1- not at all confident, 2- slightly confident, 3- moderately confident, 4 - fairly confident, 5 – very confident).

3.) Run: _____

How confident are you that you attacked the appropriate high and low priority targets?
Please rank your decision between 1 and 5 with (1- not at all confident, 2- slightly confident, 3- moderately confident, 4 - fairly confident, 5 – very confident).

4.) Run: _____

How confident are you that you attacked the appropriate high and low priority targets?
Please rank your decision between 1 and 5 with (1- not at all confident, 2- slightly confident, 3- moderately confident, 4 - fairly confident, 5 – very confident).

Sixteen WASM Condition

Participant:_____
Date: _____

1.) Run: _____

How confident are you that you attacked the appropriate high and low priority targets?
Please rank your decision between 1 and 5 with (1- not at all confident, 2- slightly confident, 3- moderately confident, 4 - fairly confident, 5 – very confident).

2.) Run: _____

How confident are you that you attacked the appropriate high and low priority targets?
Please rank your decision between 1 and 5 with (1- not at all confident, 2- slightly confident, 3- moderately confident, 4 - fairly confident, 5 – very confident).

3.) Run: _____

How confident are you that you attacked the appropriate high and low priority targets?
Please rank your decision between 1 and 5 with (1- not at all confident, 2- slightly confident, 3- moderately confident, 4 - fairly confident, 5 – very confident).

4.) Run: _____

How confident are you that you attacked the appropriate high and low priority targets?
Please rank your decision between 1 and 5 with (1- not at all confident, 2- slightly confident, 3- moderately confident, 4 - fairly confident, 5 – very confident).