



TECHNICAL REPORT FCDD-AMT-20-09

SYNERGISTIC UNMANNED MANNED INTELLIGENT TEAMING (SUMIT) FINAL REPORT

Thomas J. Alicia

Technology Development Directorate
Combat Capabilities Development Command
Aviation & Missile Center

And

Brian T. Hall

USfalcon, Inc.
100 Regency Forest Drive, Suite 150
Cary, NC 27518

And

Mikhail Terman

Perceptronic Solutions, Inc.
3141 Fairview Park Drive, Suite 415
Falls Church, VA 22042

November 2020

Distribution Statement A: Approved for public release; distribution is unlimited.

DESTRUCTION NOTICE

FOR CLASSIFIED DOCUMENTS, FOLLOW THE PROCEDURES IN DoD 5200.22-M, INDUSTRIAL SECURITY MANUAL, SECTION II-19 OR DoD 5200.1-R, INFORMATION SECURITY PROGRAM REGULATION, CHAPTER IX. FOR UNCLASSIFIED, LIMITED DOCUMENTS, DESTROY BY ANY METHOD THAT WILL PREVENT DISCLOSURE OF CONTENTS OR RECONSTRUCTION OF THE DOCUMENT.

DISCLAIMER

THE FINDINGS IN THIS REPORT ARE NOT TO BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION UNLESS SO DESIGNATED BY OTHER AUTHORIZED DOCUMENTS.

TRADE NAMES

USE OF TRADE NAMES OR MANUFACTURERS IN THIS REPORT DOES NOT CONSTITUTE AN OFFICIAL ENDORSEMENT OR APPROVAL OF THE USE OF SUCH COMMERCIAL HARDWARE OR SOFTWARE.

REPORT DOCUMENTATION PAGE				Form Approved OMB No. 0704-0188	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing 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 the burden, to the Department of Defense, Executive Service Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.					
1. REPORT DATE (DD-MM-YYYY) November 2020		2. REPORT TYPE Final		3. DATES COVERED (From - To)	
4. TITLE AND SUBTITLE Synergistic Unmanned-Manned Intelligent Teaming (SUMIT) Final Report				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) Thomas J. Alicia, Brian T. Hall, and Mikhail Terman				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Commander, U.S. Army Combat Capabilities Development Command ATTN: FCDD-AMT-AMH Redstone Arsenal, AL 35898-5000				8. PERFORMING ORGANIZATION REPORT NUMBER TR-FCDD-AMT-20-09	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Distribution Statement A: Approved for public release; distribution is unlimited.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT The Manned-Unmanned Teaming (MUM-T) employment concept is a force multiplier for United States (U.S.) Army Aviation. This collaborative employment of manned and unmanned platforms generates enhanced situation awareness, greater lethality, and improved survivability. The Synergistic Unmanned Manned Intelligent Teaming (SUMIT) program was a 6.3 Science and Technology (S&T) program evaluating the impact of innovative autonomy, decision aiding, and Human Machine Interface (HMI) technologies on an Air Mission Commander (AMC) during the execution of MUM-T scenarios. The primary purpose of SUMIT was to identify Manned-Unmanned Teaming (MUM-T) system capabilities that increase combat effectiveness and provide recommendations on metrics to assess these capabilities.					
15. SUBJECT TERMS Manned-Unmanned Teaming (MUM-T), Human Machine Interface (HMI), Autonomy, Decision Aids					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Thomas J. Alicia
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UU	92	19b. TELEPHONE NUMBER (Include area code)

ABSTRACT (CONCLUDED)

The U.S. Army Combat Capabilities Development Command (DEVCOM) Aviation & Missile Center (AvMC) recruited 40 volunteer participants with relevant experience to complete challenging MUM-T scenario vignettes in reconfigurable rotary wing flight simulators. Over the course of five system evaluations utilizing integrated vendor proposed design solutions, the results identified MUM-T system capabilities that increase combat effectiveness as well as methodological recommendations for future MUM-T technology assessment efforts.

ACKNOWLEDGMENTS

The SUMIT effort would not have been possible without the contributions and support of many people. The authors would like to express special appreciation and gratitude to Colonel Steven Braddom and Mr. Nate Bordick for their continuous support for this program; Dr. Paul Schutte, Dr. Jay Moorthy, and Mr. George Logan for starting the program down the path to success; Dr. Grant Taylor and Mr. Terry Turpin for creating the Future Open Rotorcraft Cockpit Environment (FORCE) simulator design and generating design requirements; Mr. Gerrit Burke and Mr. Tim Flynn for leading the development and management of the FORCE simulators; Mr. Frank Largo, Mr. Dan Spearing, and Mr. Patrick Ballard for building the FORCE simulators; Mr. Jason York for facility, infrastructure, and on-site personnel support; Mr. Adam Frederick for on-site personnel management; Mr. Jackson Holmes and Mr. Jermaine Davis for facilitating the integrations and evaluations and being the first responders for on-site troubleshooting; and Mr. Terry Burks for developing and integrating the Teaming Data Management (TDM) to facilitate communication between the front-end vendor interfaces and FORCE.

Additionally, our sincere gratitude to Mr. Eric Tobias and Mr. Anthony Gong for creating the revised high-speed flight model; Mr. Joe Gunderson, Mr. Nathan Mielcarek and Mr. Weiliang Dai for creating the autopilot and aided target recognition components; Mr. Tom Hart and Mr. Eric Shickel for integrating the autopilot and aided target recognition components in FORCE; Mr. Dave Durbin, Mr. Ellis Chatterton, and Mr. Jamison Hicks for evaluation support; Mr. John Martin for serving as the Subject Matter Expert (SME) Observer for every participant, and Mr. Terry Turpin for transcribing and consolidating all Air Mission Commander (AMC) interview recordings.

We also received invaluable SME support and input from Mr. Dave Bearden; Mr. Bill Huff, III; Mr. Bill Huff, IV; Mr. Adam Frederick; Mr. Terry Turpin; Mr. Matt Steele; and Mr. John Martin. Finally, this program would not have been successful without participation from the teams at Infoscitex, Toyon Research Corp., General Electric (GE) Aviation, and Perceptronics Solutions. The authors wish to thank you all for your incredible efforts and teamwork.

EXECUTIVE SUMMARY

The Manned-Unmanned Teaming (MUM-T) employment concept is a force multiplier for United States (U.S.) Army Aviation. This collaborative employment of manned and unmanned platforms generates enhanced situation awareness, greater lethality, and improved survivability. The Synergistic Unmanned Manned Intelligent Teaming (SUMIT) program was a 6.3 Science and Technology (S&T) program evaluating the impact of innovative autonomy, decision aiding, and Human Machine Interface (HMI) technologies on an Air Mission Commander (AMC) during the execution of MUM-T scenarios. The primary purpose of SUMIT was to identify capabilities for a MUM-T system that would increase combat effectiveness and provide recommendations on metrics to assess these capabilities.

The U.S. Army Combat Capabilities Development Command (DEVCOM) Aviation & Missile Center (AvMC) recruited 40 volunteer participants with relevant experience to complete challenging MUM-T scenario vignettes in reconfigurable rotary wing flight simulators. Over the course of five system evaluations utilizing integrated vendor proposed design solutions, the results indicated many impactful capabilities that merit inclusion and further assessment in future Army MUM-T systems. These capabilities include aided target recognition, cooperative engagement, high-level tasking, task sharing, task queueing, information at the point of need, automation transparency, sensor scan history and current scan status, notification of task completion, immediate direct control availability, and the ability to quickly react to emergent threats. This report describes the call for innovative MUM-T design solutions, the methods used to evaluate them, and the resulting conclusions which identify MUM-T system capabilities that increase combat effectiveness as well as methodological recommendations for future MUM-T system assessment efforts.

TABLE OF CONTENTS

	<u>Page</u>
I. INTRODUCTION	1
A. Background	1
B. Current Manned-Unmanned Teaming Limitations.....	1
C. Synergistic Unmanned-Manned Intelligent Teaming Purpose	1
II. TECHNOLOGY ASSESSMENT APPROACH	2
A. Solicitation Overview	2
B. Future Open Rotorcraft Cockpit Environment Simulators	4
C. Air Mission Commander Role.....	7
D. Scenario Vignettes	7
E. Evaluation Round Differences	7
F. Participants	10
G Assessment	10
H. Procedure	14
I. Limitations	16
III. RESULTS	17
A. Scenario Realism.....	17
B. Participant Feedback	18
C. Subject Matter Expert Observer Ratings	21
D. Situation Awareness	26
E. Workload.....	28
F. Interactions Between Situation Awareness and Workload	30
G. Automation Interventions.....	31
H. Trust and Usability	32
I. Age.....	35
IV. DISCUSSION.....	37
A. Holistic Data Interpretation	37
B. Design Feedback	39
C. Air-Launched Effects	41
D. Degraded Communications	42
V. CONCLUSIONS.....	43
REFERENCES	45
LIST OF ABBREVIATIONS, ACRONYMS, AND SYMBOLS	46

TABLE OF CONTENTS (CONCLUDED)

	<u>Page</u>
APPENDIX A: NATIONAL AERONAUTICS AND SPACE ADMINISTRATION TASK LOAD INDEX.....	A-1
APPENDIX B: BEDFORD WORKLOAD SCALE	B-1
APPENDIX C: SITUATION AWARENESS RATING TECHNIQUE	C-1
APPENDIX D. POST-STUDY SYSTEM USABILITY QUESTIONNAIRE	D-1
APPENDIX E. CONDENSED 14-ITEM SCHAEFER TRUST IN AUTOMATION SUBSCALE	E-1
APPENDIX F. GENERAL TRUST IN AUTOMATION INVENTORY	F-1
APPENDIX G. DEMOGRAPHICS SURVEY.....	G-1
APPENDIX H. POST-VIGNETTE QUESTIONNAIRE.....	H-1
APPENDIX I. SYNERGISTIC UNMANNED MANNED INTELLIGENT TEAMING PILOT INTERVIEW FEEDBACK	I-1

LIST OF ILLUSTRATIONS

<u>Figure</u>	<u>Title</u>	<u>Page</u>
1.	Representative Screenshots of Infoscitex Corporation's TECUMSA Software Suite.....	3
2.	Representative Screenshot of Toyon Research Corporation's DAM Software Suite.....	3
3.	Representative Screenshot of GE Aviation's MOMS Software Suite.....	4
4.	FORCE Simulator	5
5.	SME Observer Seated to Left of Participant During Evaluation.....	12
6.	MU Scores by Principal Task, System A	24
7.	MU Scores by Principal Task, System B	25
8.	PCA Examining Workload Factors	29
9.	Trust in Automation Responses.....	33
10.	Trust in Automation Posttest and Usability Correlation Plot	34
11.	DECIDE MU Scores by Task and Split by Age for One System.....	36

LIST OF TABLES

<u>Table</u>	<u>Title</u>	<u>Page</u>
1.	Simulation Differences by Round	8
2.	SPAM Query Prompts	13
3.	MU DECIDE Correlations	22
4.	Workload Self-Report Ratings for Round 2 Vendor System	30
5.	Workload Correlations With SART Demand Factors.....	31
6.	Workload Correlations With SART Supply Factors	31
7.	Workload Correlations With SART Understanding Factors and Overall SART Score	31
8.	Selection of Performance Metrics for One System Evaluation Across All Vignettes	32
9.	Target Resolution Metrics for One Evaluation by Age.....	35

I. INTRODUCTION

A. Background

Manned-Unmanned Teaming (MUM-T) is a critical capability and a force multiplier for United States (U.S.) Army Aviation. Incorporating a MUM-T capability—defined as the collaborative employment of manned and unmanned platforms—drives enhanced Situation Awareness (SA), greater lethality, and improved survivability. Combining the inherent strengths of manned and unmanned platforms produces synergy and overmatch with asymmetric advantages [1]. The U.S. Army MUM-T systems accomplish this by enabling a rotary wing aviator to leverage an Unmanned Aerial System (UAS), providing increased battlespace awareness and range of effects. This additional knowledge comes in the form of known threat positions and flight hazards along the flight route, known security of the battle position, Line of Sight (LOS) to targets from the firing position, and known positions of friendly troops, threats, and civilians in the target area. Additionally, the manned platform is able to perform cooperative engagements to prosecute targets without exposure and gains a top-down full picture of battle damage and weapons effects. The primary challenge is to provide this increased amount of information and control to an operator, such as an Air Mission Commander (AMC), without overloading or distracting them from their already incredibly demanding job on the battlefield. With proper design founded on solid human factors and human systems integration principles, an integrated MUM-T capability can increase SA and mission effectiveness while decreasing risk to human lives.

B. Current Manned-Unmanned Teaming Limitations

Current MUM-T systems, such as that in the AH-64E Apache Copilot Gunner's (CPG's) station, have significant limitations. As designed, the AH-64E MUM-T system can effectively operate at Level of Interoperability (LOI) 3, which allows the CPG to receive and control sensor payload video from a single UAS, though the control interface imposes high levels of workload due to the need for continuous manual control of the sensor and continuous visual information processing from the CPG. Although the AH-64E MUM-T system can provide LOI 4 control, which allows for command and control of the vehicle platform with the exception of takeoff and landing, the further increase in cognitive workload to maintain awareness of the vehicle's location, orientation, and surroundings results in Army MUM-T users rarely utilizing LOI 4 UAS control during combat. Additional limitations include the reliance on dozens of bezel buttons requiring the CPG to retain in memory the hierarchical structure of the bezel button menus three levels deep, a complex link-up procedure to cycle between UAS assets, and no visual feedback on what areas have been searched. As the Army develops the Future Vertical Lift (FVL) next generation aviation platform, the opportunity is ripe for an improved MUM-T interface design that can overcome these limitations to provide an intuitive, multi-vehicle command and control system.

C. Synergistic Unmanned Manned Intelligent Teaming Purpose

The U.S. Army Combat Capabilities Development Command (DEVCOM) Aviation & Missile Center (AvMC) designed and executed the Synergistic Unmanned Manned Intelligent Teaming (SUMIT) program with two primary purposes in mind: determine which capabilities

will increase combat effectiveness of a future MUM-T system and identify and recommend metrics to assess this impact.

A secondary purpose of SUMIT was to develop and demonstrate a Modular Open Systems Architecture (MOSA) research flight simulator that could greatly reduce the time necessary to integrate third party software and hardware design concepts for subsequent test, demonstration, and evaluation. The results of this effort are the building blocks for future MUM-T technology systems. These building blocks will enable the Army to make well-informed decisions regarding technology investment, research, and development.

II. TECHNOLOGY ASSESSMENT APPROACH

A. Solicitation Overview

The SUMIT program began with a Broad Agency Announcement (BAA) that solicited innovative MUM-T design solutions to support two rounds of evaluation. The stated objective of the SUMIT BAA was to evaluate the impact of new and recent autonomy, decision aiding, and Human Machine Interface (HMI) technologies in the execution of U.S. Army Aviation missions that utilize MUM-T capabilities. The overarching SUMIT program approach was to bring in experienced U.S. Army aviators to conduct realistic MUM-T missions in a government-owned reconfigurable research flight simulator. By integrating and performing exploratory evaluations on multiple innovative MUM-T design solutions, the SUMIT program sought to inform U.S. Army Aviation on how to evaluate novel MUM-T systems for future platforms. To do this, the government sought vendor design solutions to perform many complex functional capabilities. A subset of capabilities sought via the SUMIT BAA are shown in the following:

- Mission Planning
- Team and collaboration management
- Oversight of real-time mission execution
- Dynamic replanning for the team in reaction to events or changes in the environment
- Allocating tasks to assets or subteams based on their performance characteristics
- Coordinating target handoff between assets
- Creating and coordinating route plans for all assets in the team
- Providing continuous status updates for team SA
- Supporting control handover for LOI 2-5
- Tracking entity position data
- Maintaining restricted airspace data
- Assisting the AMC with target engagement firing solutions

After subsequent proposal evaluation and selection, three vendors received a contract award to develop their MUM-T software design solutions and integrate them into the government-owned Future Open Rotorcraft Cockpit Environment (FORCE) fixed-base research flight simulator, as shown in Section II.B. The FORCE simulators are reconfigurable and enable evaluation of innovative MUM-T design solutions and capabilities. The FORCE team and vendors worked together to rapidly integrate novel MUM-T systems for evaluation. Infoscitex Corporation developed the Tasking and Execution of Collaborative Unmanned and Manned Systems with Autonomy (TECUMSA) system, as shown in Figure 1. Toyon Research Corporation developed the Decentralized Asset Manager (DAM), as shown in Figure 2. General Electric (GE) Aviation developed the Mission Operation Management System (MOMS), as shown in Figure 3.

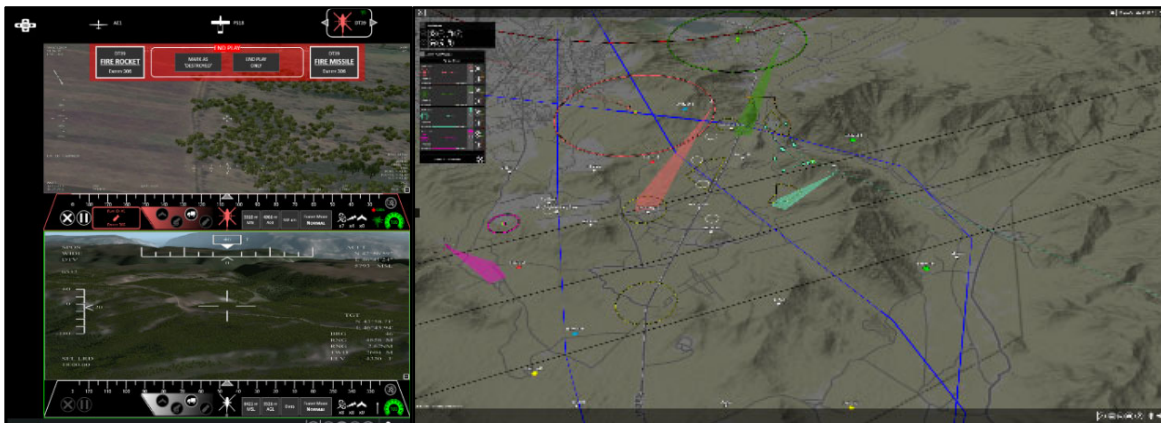


Figure 1. Representative Screenshots of Infoscitex Corporation's TECUMSA Software Suite



Figure 2. Representative Screenshot of Toyon Research Corporation's DAM Software Suite



Figure 3. Representative Screenshot of GE Aviation's MOMS Software Suite

After integrating these systems within the FORCE simulator, the SUMIT team invited experimental test pilots and retired and active duty U.S. Army Aviators with relevant operational experience to conduct a series of mission vignettes to assist the government evaluation of the impact of MUM-T capabilities on the role of an AMC. These evaluations allowed for Soldier touch points with a Science and Technology (S&T) effort that will directly influence future Warfighter effectiveness. The SUMIT program was not a competitive downselect to identify which design solution was optimal. Rather, an exploratory evaluation approach was deemed more appropriate to meet the SUMIT goals of identifying impactful capabilities and metrics. As a result, the primary conclusions from this study were derived from a synthesis of pilot feedback, Subject Matter Expert (SME) and evaluator observations, and defined, traceable metrics.

B. Future Open Rotorcraft Cockpit Environment Simulators

The FORCE simulators refer to two fixed-base research flight simulators designed, developed, fabricated, and assembled by a heterogeneous and geographically dispersed team of government civilians and contractors, as shown in Figure 4. The two simulators are identical and operate independently to increase the capacity to conduct technology evaluations in parallel. Located in Huntsville, AL, each of the two simulation environments consists of two primary system components: cockpit with side-by-side seating arrangement and Simulation Management Station (SMS).



a. Cockpit



b. SMS

Figure 4. FORCE Simulator

FORCE is a government-owned and purpose-built MOSA demonstration environment designed to allow integration and evaluation of innovative MUM-T design solutions. SUMIT vendors integrated their systems into FORCE in order to demonstrate their technologies within a representative mission environment. FORCE is built on Future Airborne Capability Environment (FACE) conformant software and provides a Teaming Data Manager (TDM) interface for other systems to integrate with FORCE. The TDM interface provides UAS state reporting and control at an abstract level designed to generalize the unique proprietary message formats typically required to communicate with disparate vehicle platforms [2]. This ensured that the FORCE simulators were flexible to multiple varied vendor interfaces and integration methods.

This approach was validated by first integrating a third-party government-owned MUM-T research system into a FORCE simulator as a functioning proof of concept prior to integrating any vendor systems. This MUM-T system, known as Supervisory Controller for Optimal Role Allocation for Cueing of Human Operators (SCORCH), was previously developed to test a novel pilot-vehicle interface and innovative autonomous UAS behaviors to determine the ability of a CPG or AMC to effectively command teams of multiple, autonomous UAS assets [3]. Data were collected in Spring 2018 from eight participants using the SCORCH system integrated in FORCE to validate the MOSA architecture and flexibility of the FORCE cockpit design and verify data collection hardware and software functionality through a complete system checkout.

The flexibility of FORCE also extends to the internal software, including the ability to update the aircraft flight models and providing support for, but not requiring, standardized UAS communication protocols, such as Standardization Agreement (STANAG) 4586 [4]. This flexibility, together with the government-furnished documentation provided in advance to vendors specifying detailed architecture and interface descriptions, resulted in rapid integration with the contractor systems. The typical multi-week or multi-month integration timeline was reduced to as little as 3 days, with no system integration taking longer than 2 weeks to complete.

The FORCE simulators are highly reconfigurable, serving as a shell in which third parties can integrate various hardware configurations. Over the course of the evaluations, the SUMIT team reconfigured the FORCE simulators numerous times to utilize varying numbers, types, and sizes of touchscreen displays as well as external tablets for chat messages and a variety of joysticks, game controllers, and other hardware.

Lastly, FORCE is portable. Three trained personnel can easily disassemble and pack the entire simulation system into a few transport containers for easy conveyance to exhibit or demonstrate as desired. The SUMIT team transported a FORCE simulator in this manner on a box truck from the evaluation building location to the Association of the U.S. Army Global Force Symposium and Exposition in Huntsville, AL, where SUMIT team members were able to unpack and assemble FORCE to functioning status in less than 3 hours from the time the containers arrived to the show floor.

C. Air Mission Commander Role

The SUMIT program assessed the impact of innovative autonomy, HMI, and decision aids on an AMC performing MUM-T tasks. The traditional role of the AMC is to be in command of all aircraft in a flight. Reference 5 states that the unit commander designates this command responsibility based on recent aviation experience, maturity, judgment, abilities for mission SA, and understanding of commander's intent. As the nature of conflict changes from predominately Counter-Insurgency (COIN) operations to potential decisive action missions, the focus and responsibility of the AMC must also adapt. For SUMIT evaluation purposes, each participant AMC was responsible for planning and replanning the MUM-T mission. During mission execution, the AMC was responsible for oversight and management of the team and individual team platforms. This included the piloted manned aircraft, known as the ownship, and each of the unmanned assets. Additional AMC responsibilities included acquisition, control, and release of various unmanned assets; control of sensor payloads; planning and execution of engagement sequences; mission monitoring and contingency handling; and reporting to and receiving directives from higher command. The AMC met these responsibilities by interacting with the teamed assets in the following way:

- Team mission command level—for example, assigning functions to multiple UAS assets
- Individual task level—for example, managing a single UAS
- Individual resource level—for example, controlling the sensor payload on a UAS as needed by mission context

D. Scenario Vignettes

A working group of SMEs created one training vignette and four test vignettes for each of the two rounds of evaluation. All scenario vignettes were based on movement to a Landing Zone (LZ) and LZ preparation incorporating a varying number and availability of UAS assets. The mission description was the same for each scenario—prepare the ingress route and LZ for a hasty air assault by scanning and clearing the ingress route and LZ of any threats or obstacles. Secondary tasks included searching other areas of interest depending on the Mission, Enemy, Terrain, Troops Available, Time, and Civilian (METT-TC) considerations and destroying anti-aircraft weapon systems. The majority of mission vignettes were 30 minutes in length, while the final vignette in Round 2 was 1 hour in duration. Standard rules of engagement applied, with the exception that if an AMC was able to positively identify a threat, they were cleared to engage without requesting permission to fire.

E. Evaluation Round Differences

Due to the exploratory nature of the SUMIT S&T program, many enhancements and modifications were made to the FORCE simulators and the scenario missions and environments in the months between the first and second rounds of evaluation, as shown in Table 1. Additional changes included vendor modifications to their design solutions hardware and software and an increased number of assessment measures.

Table 1. Simulation Differences by Round

	Round 1	Round 2
Mission	COIN Operations	Decisive attack against a prepared, coordinated defense
Environment	Khairabad, Afghanistan	Caucasus Mountains
Ownship	Black Hawk-like Aircraft	FVL-like Aircraft
Weapons	Hellfire-R Missiles	Hellfire-R missiles, APKWS Rockets, LRPF
Available UAS Assets	Gray Eagle, Shadow, Future Scout	Gray Eagle, Shadow, Fire-X, ALTIUS ALE
Number of Sensor Feeds	Five Simultaneous Sensor Feeds	Ownship sensor plus one other live sensor video at a time
RWRs	N/A	Available
Shot Detection	TOC Voice Reports When Receiving Fire (Direction Only)	Acoustic detection and radar detection of incoming fire with point of origin through TDM and text messaging

1. Mission Type and Environment

The first round of data collection with the three vendors occurred with one FORCE simulator between September 2018 and May 2019, with scenario vignettes based on COIN operations and suppression of terrorist activity in Afghanistan. Five months after concluding Round 1 evaluations, Round 2 evaluations commenced with new vignettes tailored for decisive attack operations against a prepared, coordinated, near-peer threat defense in the forested Decisive Action Training Environment (DATE) region near the Caucasus Mountains in Eastern Europe. This shift in context reflected the mission types, threat, and operating environment most likely to utilize the FVL platform. These Round 2 evaluations with two vendors occurred between October 2019 and February 2020. All scenario vignettes in both rounds retained the same primary tasks to scan and clear the ingress route and primary LZ in support of a hasty combat assault.

2. Ownship Configuration

Other modifications between rounds included the ownship configuration and flight model. In Round 1, the ownship was a generic representation of a Black Hawk aircraft with a right seat pilot and a maximum speed of approximately 190 knots (kn). In Round 2, the ownship was reconfigured with a high-speed flight model to generically represent envisioned FVL flight speeds upwards of 250 kn. Furthermore, a limited autopilot capability came online during Round 2, enabling an exploration of a single pilot in command to who is able to navigate and control the ownship while also operating the MUM-T system. The second FORCE simulator also became operational in this time between rounds, enabling more efficient data collection by allowing one simulator to be used for participant data collection while a second participant was undergoing system training in the second simulator.

3. Available Weapons and Assets

The participant's available weapons options were enhanced between rounds to be more representative of an operational environment. In Round 1, Hellfire-R missiles were the only available weapons. In Round 2, the ownship and select UAS assets were upgraded to include Hellfire-R missiles and Advanced Precision Kill Weapon System (APKWS) rockets. Additionally, Round 2 evaluations included the ability to call for Long-Range Precision Fire (LRPF) from a remote location.

The number and type of UAS assets available changed over the course of the SUMIT evaluations. In the first round, the AMC was provided with a Shadow, Gray Eagle, and notional Future Scout, which is an unmanned rotary wing platform armed with a laser designator and Hellfire-R missiles. The specific availability of assets varied by vignette. For Round Two, again the availability of unmanned assets was varied by vignette. The SUMIT team provided the AMC with a Shadow; Gray Eagle; notional Fire-X platform, which is modeled on an MQ-8 Fire Scout but armed with a laser designator, Hellfire-R missiles and APKWS rockets; and four Air-Launched Effects (ALEs) modeled on the Area-I Air-Launched, Tube-Integrated, Unmanned System (ALTIUS) platform. The ALE program is a separate development effort that is launching unmanned vehicles from manned and eventually unmanned platforms. In the second round of SUMIT evaluations, two ALEs were carried on board the AMC's ownship, and two ALEs were carried on the Gray Eagle. They provided insight into the thought process, possible use, and strategy from the AMC with regard to this emerging technology. Evaluators intentionally provided no guidance on strategies to utilize the ALEs in order to explore how they might be used.

4. Number of Sensor Feeds

In the first round of evaluation, the government did not specify a maximum number of live video sensor feeds displayed on screen. This technological assumption was based on the notion of multiple high bandwidth video feeds displayed to a MUM-T operator. As a result, vendor systems were allowed to display all UAS sensor payloads simultaneously. Prior to the second round of evaluation, this technological assumption was revised to explore the impact on design solutions when presented with more limited data link connections. This first step of generalizing a degraded communications environment added a constraint for the vendor systems to display only the ownship sensor and a maximum of one additional sensor video feed at a time. This required the vendor to incorporate an ability for the participant to switch from one sensor to another as desired or necessary.

5. Radar and Shot Detection

In Round 1, the simulator did not provide Radar Warning Receiver (RWR) messages to the vendor system. The nature of the COIN threats in the vignettes did not justify the inclusion of such a capability requirement. As the missions changed for Round 2 to a decisive action against near-peer threats, the justification was realized and an RWR capability was added to the FORCE simulators.

Additionally, the level of detail informing a participant when their ownship or UAS assets were fired upon in Round 1 was limited to the Tactical Operations Center (TOC) communicating over the radio which asset was taking fire and from which direction. The participant could interpret from their mission briefing the effective threat weapon range of 2.5 kilometers (km) and approximate a location to search based on the direction given from the TOC; however, this was an approximation. In Round 2, this level of detail was increased to represent acoustic detection of heavy gunfire and radar detection of missile and rocket fire with point of origin messages communicated to the vendor system to subsequently display to the participant.

F. Participants

Forty participants were recruited from the U.S. Army Aviation community. This allowed for eight participants each for five total vendor evaluations, including three vendors in Round 1 and two vendors in Round 2. Participants in the first round were current and former U.S. Army Aviators, some with experience as experimental test pilots. Seeking to maximize opportunities to provide the aforementioned Soldier touch points with cutting edge S&T programs, participants in the second round of evaluations were active duty U.S. Army Aviators from Fort Campbell, KY. Of the 40 participants, one participant was dismissed after significant technical difficulties with the simulator, resulting in valid data for 39 participants. The sample included 37 males and 2 females ranging in age from 26 to 60 years ($M = 40.6$; $SD = 7.6$). All participants in the sample reported that they had rotary wing experience, ranging from 295 to 8,000 hours ($M = 2,739.0$; $SD = 1,832.3$), while 36 out of the 39 participants had combat flight hours ($M = 934.1$; $SD = 821.0$). Twenty of the 39 participants had previously teamed with unmanned systems in training, and 14 had teamed with unmanned systems in combat. Of those 14 with combat MUM-T experience, 13 had served as an AMC during those missions.

G. Assessment

Evaluators assessed the capabilities of the MUM-T systems by examining participant feedback, SME and evaluator observations, and defining traceable metrics. The vast amounts of subjective and objective data collected were interpreted with the intent to provide observations and lessons learned on MUM-T capabilities and metrics, suggestions for further exploration and recommendations on the impact of advanced autonomy, HMI, and decision aids on an AMC performing MUM-T missions. The exploratory, nonexperimental nature of the SUMIT program allowed for lessons learned to be identified and applied between vendors and between the two rounds of evaluation. This encouraged a number of modifications to be introduced in order to maximize the value of the data collection opportunities. These changed elements included vendor system enhancements, scenario mission and environment changes, evolving system requirements and constraints, changes in recruited participant demographics, additional FORCE functionality, additional metrics, questionnaires, and variables, and revised SME Observer subtasks and ratings.

1. Participant Feedback

Evaluators elicited participant feedback through a series of short post-vignette questionnaires that were focused on the most difficult and challenging aspects of the vignettes, mistakes a novice or inexperienced AMC might make, important capabilities present or absent in the system, and the contribution of the MUM-T capability to mission success. After each participant completed the fourth and final vignette and questionnaire, the evaluator administered an additional comprehensive end-of-evaluation interview. This extensive semi-structured interview inquired about the participant's opinions on the autonomy, HMI, and decision aids, often leading to additional follow-up questions and discussions regarding specifics of the participant's performance and observations by the evaluator and SME Observer. Additionally, the interview contained a series of open-ended questions asking about topics, such as controlling multiple UASs simultaneously, workload, SA, trust in automation, future technology systems, and the role of the AMC. Finally, the interview also touched on the simulation evaluation, participant training, vignettes, FORCE simulator, and simulation evaluation experience. This knowledge elicitation approach resulted in over 100 hours of audio recordings from the evaluation participants that were subsequently transcribed and evaluated.

2. SME and Evaluator Observations

The SUMIT program also utilized a contractor-developed system known as Diagnostic Engine for Cognitive Improvement and Decision Effectiveness (DECIDE) that is installed on a Microsoft Surface tablet. Perceptronics Solutions originally developed the DECIDE tool for Naval Sea Systems Command (NAVSEA) and adapted the system for SUMIT to provide insight regarding the contribution of technologies to the successful accomplishment of mission tasks. An SME working group collaborated with Perceptronics Solutions to perform a functional task decomposition of each task included in the mission vignettes. This working group decomposed the scenario events to extract operational tasks and low-level mission functions. Developers relied on SME input to weight the importance of each of the compressed operational tasks to completion of the overall scenario task. The functional task decomposition and weighting was performed again alongside other DECIDE enhancements prior to Round 2 evaluations to align with the updated mission focus, improve DECIDE functionality, and provide greater granularity for SME Observer input by allowing a selection of questions to be answered on a five-point scale as opposed to a binary yes or no.

The SME Observer sat to the side and just behind the participant during the evaluations, as shown in Figure 5, and identified and selected on the DECIDE tablet an operational task being performed by the participant. The DECIDE system then selected and presented the SME with specific questions about the level of system support for the participant to perform those tasks. These questions were identified as more likely to be highly diagnostic, reducing the number of questions that the DECIDE observer had to answer by prioritizing questions most likely to be informative. The questions presented were traceable to three predetermined categories, which include Requirements Satisfaction, Military Utility (MU), and Usability. Participant interactions with the systems were monitored by an evaluator and recorded on audio and video for follow-up and post hoc analysis. Additionally, the evaluator consistently prompted participants to verbalize their decision making and thought processes as much as possible during scenario execution, and these comments were recorded as well. This enabled

the evaluator to observe traits of high performers, instances of expert decision making and prioritization, and features and capabilities that led to a successful action or decision, including participant frustration with a less than optimal result that may not otherwise be represented in the data.



Figure 5. SME Observer Seated to Left of Participant During Evaluation

3. Measures

The SUMIT program utilized a vast array of measures to inform further exploration. One such measure was the Level of Automation (LOA) provided by the vendor systems for a predetermined list of capabilities. Each vendor assigned the listed capabilities with an LOA ranging from zero to three, such as classify objects, track moving targets, store targets for future reference, and modify UAS behavior based on results of reconnaissance. Assigning a capability a level zero indicated that capability was not supported by the system. Assigning LOA 1 indicated system support for manual control, which is human-in-the-loop. LOA 2 indicated supervisory control, which is human-on-the-loop, while LOA 3 indicated a fully or highly automated capability, which is human-out-of-the-loop.

Objective system-based measures indicated primary mission effectiveness. Evaluators instructed all participants that their primary task was to scan and clear the ingress route and the primary LZ. This was measured by the fraction of the ingress route and LZ scanned by a sensor. Other secondary tasks became relevant throughout mission execution, such as scanning additional areas of interest, locating a high value target, or destroying anti-aircraft weapon systems. Relevant metrics for secondary task performance included the fraction of total areas of interest covered by a sensor and the number of anti-aircraft weapon systems destroyed.

Other system-based measures included the total and percentage of entities at each step of the workflow for resolving detected targets, including detect, classify, track, engage, and destroy; time to complete each step; classification accuracy; time to complete area and route reconnaissance tasks; asset location; ALE launch times; sensor utilization efficiency; number of times a participant interrupted the automation; duration of automation interruptions; duration of friendly assets vulnerable to threats; and number of friendly assets shot down.

In addition to the volumes of system-based data, there were also several subjective questionnaire-based metrics. Measures indicating workload included the National Aeronautics and Space Administration (NASA) Task Load Index (TLX) questionnaire, as shown in Appendix A, and Bedford Workload Scale questionnaire, as shown in Appendix B. SA was subjectively assessed using multiple techniques. Both the participant and SME Observer rated participant SA by completing the Situation Awareness Rating Technique (SART) questionnaire, as shown in Appendix C. The second round of evaluations also incorporated the Situation Present Assessment Method (SPAM) [6] by providing queries to the participant over the radio during mission execution. The SPAM technique relies on the concept that SA assessment should allow for the information the participant has available to them in an interface. Each query prompt would begin at a preset time with the evaluator asking over the headset “AMC, are you ready for a query?” Participants were instructed that the SPAM queries were not to take priority over the mission tasks and to wait until they were ready before acknowledging as such. Once the participant confirmed that they were ready, the experimenter provided the query over the headset. The queries posed in these evaluations, their timing, and the intended level of SA [7] are shown in Table 2.

Table 2. SPAM Query Prompts

Query	Prompt Time	SA Level
In 5 minutes, will ownship be east or west of primary route?	8:00	3-Projection
How many units under your control are north of the FLOT?	12:00	1-Perception
Have you detected any large concentrations of civilians in TAI 63?	20:00	2-Comprehension
What is the closest UAV to your current position?	25:00	2-Comprehension

The Post-Study System Usability Questionnaire (PSSUQ), as shown in Appendix D, was selected for the usability assessment. Trust in automation was assessed by administering a pretest and posttest format of the condensed Schaefer Trust in Automation Scale, as shown in Appendix E, as well as a General Trust in Automation Inventory questionnaire in Round 2, as shown in Appendix F.

Finally, Round 2 also saw the inclusion of Heart Rate Variability (HRV) as an additional potential indicator for mental workload. HRV can be measured through many variables from the time and frequency domains. A literature review of HRV measures found that a number of these measures were significantly related to workload across multiple studies [8]. One of the most consistently proven HRV metrics for measuring workload appears

to be the ratio between the power in the Low-Frequency (LF) region of the frequency domain (0.04-0.15 hertz (Hz)) and the power in the High-Frequency (HF) region of the frequency domain (0.15-0.4 Hz). This LF and HF ratio is thought to be associated with workload because it tracks the activation of the Sympathetic Nervous System (SNS) and Parasympathetic Nervous Systems (PNS). Indeed, Reference 9 says that “The assumptions underlying the LF/HF ratio is that LF power may be generated by the SNS while HF power is produced by the PNS. In this model, a low LF/HF ratio reflects parasympathetic dominance...In contrast, a high LF/HF ratio indicates sympathetic dominance.” These assumptions are supported by the fact that the LF and HF ratio was found to be significantly positively associated with workload in 12 out of 16 studies that were reviewed in Reference 8. Thus, the LF and HF ratio tends to increase as the workload of the individual grows, presumably because this additional workload produces greater activity in the sympathetic nervous system.

H. Procedure

The SUMIT evaluation consisted of one training day and one evaluation day for each participant. Upon arrival, the evaluator asked the participant to sign an informed consent document consenting to audio and video recording. The participant completed a demographics questionnaire, as shown in Appendix G, followed by the evaluator providing a program overview describing the SUMIT program and purpose. The participant then received classroom training from a vendor trainer that consisted of PowerPoint slides depicting the system that they would be using. This classroom training was only intended to provide a first look at the interface and the features available to the participant and would take no more than 20 minutes in order to move the participant to the simulator for hands-on training as soon as possible. Once complete with classroom training, the participant completed the General Trust in Automation questionnaire, which was added prior to Round 2, and the more specific Schaefer Trust in Automation Scale. Then, the participant was moved to the FORCE simulator to begin their hands-on training. This training with the participant in the cockpit would range from 2 to 6 hours depending on the instruction, including periodic breaks. The vendor trainer guided the participant through a structured training checklist, describing all of the features in the system and how to use them. The instructions were delivered in a consistent manner for each participant within a system evaluation. Importantly, the trainers were instructed not to provide any suggestion of tactics or strategy but to only provide information on the available features and how to use them. The training session was performed with an instructional vignette that included elements representative of events that the participant would see in the operational vignettes the following day.

Upon completion of the instructor-led hands-on training session and after ensuring the participant was comfortable and proficient with the system, the participant moved out of the cockpit to a briefing room to receive the mission briefing for the training vignette. Participants were provided an abbreviated military operations order and mission graphics. They were allowed up to 20 minutes to plan their mission. Once complete and with any questions answered, the participant returned to the FORCE cockpit with their plan and any desired materials. The evaluation team would begin audio and video recording, and the evaluator asked the participant to verbally brief their mission plan over the headset to the right seat pilot or TOC in the case of an autopilot system with no right seat pilot. This was done primarily to inform the evaluation team of the participant’s intent going into the vignette for later interpretation of

their initial actions and decisions. Once complete, the evaluation team would activate the prerecorded rotor noise, and the participant would begin an untimed training vignette with the vendor trainer sitting next to them able to identify and provide direct feedback on misuse or disuse of system features they had been trained on. These untimed training scenarios would usually last 30-45 minutes, concluding when the participant had performed every mission task element at least once.

After this untimed training vignette, the participant would receive a break while the simulator was reset and any last comments on the usage of the system were conveyed. Then, the participant would perform the same vignette training a second time without trainer support to set the participant expectations and replicate the conditions of what they would experience in the actual test vignettes. There was no real-time trainer feedback, the room lights were turned off, recordings of rotor noise and background radio chatter were turned on, and the vignette was stopped after 30 minutes. The background chatter was prerecorded, realistic, operational chatter intended to represent the types of communication an AMC would hear in the operational environment. At the 30 minute mark, the TOC asked the AMC over the headset if they recommended a go or no-go for the incoming air assault, which was based on whether they had cleared the ingress route and LZ. After completing this timed training vignette, the vendor trainer, who had been observing through a monitor in a separate room, was allowed to provide instructional feedback or reminders of system features that could have been used without providing strategy.

Once the timed training vignette was complete, the participant moved back to the briefing room where the evaluator administered the Bedford, NASA TLX, and SART questionnaires for participant practice and familiarity. The participant was encouraged to ask questions to clarify any confusion about the questionnaires or term definitions. The evaluator then turned on the audio recorder, and the participant verbally completed the post-vignette interview questionnaire, as shown in Appendix H. Finally, the evaluator collected all training materials that the participant may have marked on and dismissed the participant for the day.

When the participant arrived on the morning of the second day, they had 30 minutes of hands-on refresher training in the FORCE simulator with the vendor trainer. Once the participant felt comfortable and had demonstrated proficiency with the system, they were moved to the briefing room to attach the heart rate monitor, included in Round 2; received mission orders; and began forming their mission plan. After completing the mission planning, the participant moved to the cockpit for the first data collection vignette. The evaluation team turned on the audio and video recording, and the AMC briefed their mission plan to their right seat pilot or the TOC over the radio. The evaluation team turned off the room lights and turned on the rotor noise and radio communication chatter, then data collection began. At the 30-minute mark or the 1-hour mark for vignettes that lasted an hour, the TOC would ask the AMC if they recommended a go or no-go for the incoming assault force, record the response, and stop the scenario. The participant would immediately move to the briefing room to complete the questionnaires and post-vignette interview questions, then take a short break and come back to the briefing room to repeat this process for the next three data collection vignettes.

Once the fourth and final vignette was finished, the participant would complete the usual questionnaires plus a posttest Schaefer Trust in Automation questionnaire and PSSUQ and begin the comprehensive end-of-evaluation interview. Afterwards, the evaluation team would answer any questions that the participant had about the scenarios or their performance, then thank and dismiss the participant.

I. Limitations

The full SUMIT mission scenarios operated via many separate simulation layers communicating with each other. These layers included the front-end vendor interface, UAS vehicle models, ownship flight model, terrain model, simulation environment on the terrain, objects and entities within the environment and their associated scripted behaviors, image generators for each sensor payload, and finally, translation layers to facilitate fluent communication between these components. The complex nature of this simulation ensured that there were periodic software glitches. This likelihood necessitated a mitigation strategy depending on the severity and timing of the issue to minimize interference and intrusiveness on the participant. Skilled technical personnel were on-site and available for real-time troubleshooting if the issue could be quickly fixed with minimal disruption to the mission, for example, restarting a crashed software component for one sensor payload. More serious or disruptive glitches, such as an HMI crash, were handled on a case-by-case basis. The vignette could be restarted from the beginning if the issue was early enough to not reveal critical events or emergent tasks. If the mission was too far along to consider starting over, the evaluation team attempted to note asset state and location at the moment of the crash and restart the vignette from as close in time and state as possible to complete the remainder of the mission. Full mission vignette restarts such as this were nonetheless a last resort and quite rare.

Furthermore, the vendor HMI capabilities were innovative MUM-T design solutions with futuristic capabilities beyond the participants' operational experiences. On occasion, participants were not clear if unexpected system behavior was due to a system error or their misuse or misunderstanding of the system. They responded to this confusion in a variety of ways. Some participants learned to recognize glitches and developed workarounds to compensate or prevent them from reoccurring, while other participants abandoned some tasks altogether due to frustration. These strategies and confusion were verbalized by the participants, monitored by the evaluator, and recorded for subsequent interpretation.

Additionally, the mission vignettes were not randomized or counterbalanced in presentation order, subjecting performance to learning effects over time as the participants continued to become more proficient with the system over the course of the four evaluation vignettes. This decision was made because the evaluation team wanted to increase the number of UAS assets across the scenarios. In Round 1 for instance, Vignettes 1 and 2 had no Future Scout asset. In Vignette 3, the participant started with the Future Scout but a scripted event in the scenario resulted in the Future Scout being shot down and unavailable, while Vignette 4 gave the participant control of all UAS assets for the entire scenario. By design, as the participant was given more assets to use, they were also given more tasks to complete. Although the vignettes did not necessarily increase in complexity, they did increase in tasks assigned because of the increased number of assets available to the participant.

Lastly, the open-ended approach to the mission vignettes, where participants were told the mission priorities but were not instructed on tactics or strategy to accomplish them, meant every participant performed the vignette differently. This intentionally ambiguous instruction resulted in some surprising strategies, such as participants choosing not to fire any weapons until all targets were identified and all assets were in firing position, then suddenly engaging all targets at once. Other participants assigned specific assets to tasks based on considerations that were not modeled in the simulation, such as noise and detectability. This ambiguity also led to some participants not demonstrating or experiencing every system capability. For instance, when using a system that had the capability to provide notifications upon automated task completion, participants often detected a threat while performing an autonomous area scan and interrupted the scan to engage the threat that prevented the automated area scan task from completing and thus never received the task complete notification. These examples also illustrated the fragility of performing statistical analysis in an environment based upon so few participants. The relationship between the capabilities and the outcome measures swings dramatically based upon the inclusion or exclusion of a single data point.

Nevertheless, there were significant attempts to inject evaluation control where possible. No questionnaires or measures were modified during a vendor system evaluation. Vendor system training was provided in a structured checklist approach to ensure that all participants were consistently instructed on all features and capabilities of a system. Though participants were not expected to become expert users within one day of training, they each demonstrated a level of proficiency deemed acceptable by the vendor and evaluation team and were asked if they felt comfortable and proficient with the system before proceeding into the first training vignette. Participants performed this training vignette twice in addition to refresher training on Day 2 prior to any data collection mission. Furthermore, once the vendor integrated their system and was ready for the first evaluation participant, they were not permitted to add or modify system features during an evaluation data collection, although they were encouraged to correct any identified issues due to software glitches.

III. RESULTS

A. Scenario Realism

The evaluation design incorporated opportunities to check the validity of the measures and data. The most straightforward type of validity is face validity, which asks if the content of the evaluation appears to be effective or suitable to its aims. To this end, participants were asked after every vignette if they thought the vignette as presented was realistic. One Round 1 participant stated that the vignettes were unrealistic only in that a current day AMC would not have command and control of multiple UAS assets and sensor payloads simultaneously. All other participants stated that the events, terrain, environment, and enemy threats were realistic for the envisioned future of MUM-T operations.

B. Participant Feedback

All pilot feedback from the post-vignette interview and the end-of-evaluation interview was recorded, transcribed, and consolidated, as shown in Appendix I, to extract patterns and contribute to determining which capabilities increased combat effectiveness. These audio recordings proved to be an excellent source of context rich information robust to any limitations present with the quantitative data. Often times, points of confusion or frustration with the system resulted in the most valuable feedback since the participant was able to describe exactly what capabilities were missing or lacking in the system that led to their frustration. Conversely, participant satisfaction and emphasis on the importance of an included system capability were able to override quantitative performance data that was impacted by glitches or other external variables. The evaluation team strove to convey the focus of the SUMIT program on the system capabilities rather than specific usability improvements, and the participants understood and supported the focus of the evaluations. The findings from the participant feedback indicate impactful capabilities that are critical for future MUM-T systems:

- Aided target detection—moving and stationary targets
- Cooperative engagement
- High-level tasking or play calling
- Task sharing, for example, conduct reconnaissance with multiple assets
- Task queueing
- Information at the point of need, for example, latitude, longitude, or Military Grid Reference System (MGRS) displayed on a map, or a compass displayed in sensor view
- Automation transparency, defined as providing accurate perception of the system's intent and performance
- Sensor scan history and current scan status
- Notification of task completion
- Immediate direct control availability for critical tasks
- Ability to quickly react to threats

Aided target detection is an essential capability for any MUM-T system. A human cannot focus on two displays at the same instant. Thus, any time spent looking at a map screen or out the window is time not spent attending to the sensor payload view, while focusing on one sensor payload results in only peripheral vision on other sensor payloads at best. This performance decrement was realized all too often in the form of targets in full view of a sensor being completely missed by the participant. While this aided target detection capability must function for both moving and stationary targets, the resulting action of the automated system may vary. Examples of resulting actions from an aided target detection system detecting a target included adding an icon to the map screen, presenting an alert to the operator, or interrupting the automated task to lock-on and track the discovered target.

Cooperative engagement is the process of using one asset to laser designate a target while a second asset engages the target. This allows the operator to fully utilize the abilities of unmanned assets while keeping the manned platform safely out of vulnerable range. The increased flexibility derived from this capability can also enable faster target prosecution by allowing small, unarmed assets to simultaneously scan multiple areas of interest without requiring them to be fully capable of target engagement and acknowledging that they can provide a laser designation for identified targets to be prosecuted by an armed teammate from a greater distance. Cooperative engagements are inherently more complex and have more stringent positioning requirements to ensure safety and effectiveness; therefore, it is important that the system also assists in the positioning of the assets involved to satisfy these constraints.

High-level tasking, sometimes known as play calling, eases the burden of multi-vehicle command and control. The metaphor of a coach calling plays applies—a coach rarely tells each individual player where they need to position themselves and how to do their task during a game. Rather, the coach calls a play and all players know their proper position and actions. By incorporating play calling into MUM-T missions, the operator can be relieved from micromanaging each asset. Instead of specifying individual waypoints for each asset, for instance, the operator can specify a higher level task, such as Recon Route Secura, and expect the asset to establish its own flight path to execute the route scan.

The next capability in this list is task sharing, which means to perform one task with multiple assets. A MUM-T system design that can split a lengthy ingress route or a large area of interest into multiple segments for instance, each scanned by a separate platform, greatly increases the efficiency of the scan by reducing idle time for untasked assets and decreasing the time necessary to complete the task.

Task queuing was also indicated by the participants to be particularly helpful for performing their MUM-T missions. Setting up a future task while an asset is currently performing a task reduces idle time of the assets when they would otherwise be awaiting another task. These tasks could be added to an active task list for an available asset to initiate as soon as its ongoing task is complete. Alternatively, the tasks could be added to a list of inactive tasks that can be quickly selected later, allowing the operator to establish a mission plan up front and quickly adapt later if necessary by modifying or canceling the queued task, rather than taking their full attention to create a new task and specify necessary parameters in the middle of a mission with many moving parts.

Another capability identified as crucial from participant feedback and evaluator observation is providing information at the point of need. This can be described as presenting relevant information where the operator needs it. This concept avoids forcing an operator to assemble pieces of information from multiple disparate areas of attentional focus.

One example of this is providing a navigation compass inside of the sensor payload view so that the operator gains SA of their orientation and potentially their location in a familiar environment without diverting attention away from the payload view and onto a map. Likewise, incorporating earth-referenced symbology within the sensor payload view, such as overlaying an earth referenced line and label along a route for a route reconnaissance, would greatly increase SA and aid asset employment decision making while reducing the number of times the operator

needs to divert attention to a different source. MUM-T system designers should strive to incorporate decision aids within the area of decision maker focus at that moment and location of need. Other examples derived from the SUMIT evaluations include providing gridlines for reference within the map display and an integrated chat capability within an existing primary information display.

An additional high-level capability derived from participant feedback was a demand for automation transparency. This concept has been presented to mean many things; however, in this case, it is defined as “the quality of an interface pertaining to its abilities to afford an operator’s comprehension about an intelligent agent’s intent, performance, future plans, and reasoning process” [10]. In practice during the SUMIT evaluations, three primary examples of automation transparency included depicting the planned flight path for all assets, presenting asset task assignments and status, and informing the user of any constraints that might prohibit task assignment, such as an unarmed asset unable to engage a target or a fast moving asset unable to make a tight turn to follow waypoints spaced too close together. The inverse of automation transparency, which is automation opacity, resulted in significant frustration for the participants and often led to them cancelling automated tasks and attempting to force the task via manual control without understanding the system reasoning that prevented task completion.

Next was the capability to see the sensor scan history and current scan status. Especially in systems with automated scan capability, participants often wanted to see where the sensors had already scanned, where they were scanning now, and what remained to be scanned. This information increased SA as well as participant confidence that the scanning tasks were proceeding according to the specified plan. Participants also indicated a desire to see how current or stale the scan history was, noting that a route that was scanned and determined clear 25 minutes earlier may no longer be clear. Enabling this functionality as an overlay that can be filtered on and off minimizes occlusion of important map elements that may otherwise be covered by the sensor scan history. Finally, showing participants the sensor scan history consistently received positive participant comments and feedback.

Relatedly, a seemingly small capability that was mentioned numerous times by participants across all of the vendor systems was notification of task completion. Especially in a multitasking setting, operators cannot be assumed to be actively monitoring the status of every asset and all tasks being performed at all times. They may have shifted or zoomed the map to where that asset is no longer in view, or they may be focused on tracking a high value target. For vendor systems that did not provide this information, assets would typically complete a task and then remain idle, loitering unnoticed by the participant who was otherwise occupied by other priorities. This inefficient use of assets frequently led to participant frustration when they realized they had an available asset that had been waiting for instruction from them for a significant period of time.

Perhaps the most discussed capability throughout the entire SUMIT evaluation was the ability for the participants to take immediate direct manual control for critical tasks, specifically target engagement and sensor management. Many times, participants had a high value target—whether a person, vehicle, or anti-aircraft weapon system—within their sensor payload view but were unable to prosecute the target for one reason or another. Often, the reason was too much automation, frequently described by the participants as the system not

letting them prosecute the target. Other times, system constraints required a user to classify a target as hostile within the interface before enabling the target engagement play. Requiring a participant to call an engagement play and let the system determine which asset it decides is most appropriate or requiring the user to set specific parameters in order to select the asset they were already using to look at the target was not conducive to high performance. A common sentiment expressed was “I’ve positively identified this target as a threat. I am staring at the target in my sensor right now—just let me shoot it.”

This desire for manual control also frequently manifested when participants wanted to make small adjustments to the sensor stare point. The control mechanism, whether using a joystick, a touchscreen, or some other physical control, did not seem to have a discernable impact as much as systems that were not able to provide this level of manual control at all. Consider the situation where a participant is monitoring an automated route reconnaissance task and noticed a suspicious vehicle on the edge of the sensor payload view that was not picked up by the aided target detection system. The participants wanted to be able to immediately interrupt the automated scan, grab control of the sensor, and walk it on to the suspicious vehicle for further investigation. This was not always possible with all iterations of the vendors’ design solutions and when absent, frequently resulted in missed opportunities and significant participant frustration.

The final capability in this list is the ability to quickly react to emergent threats. In these mission vignettes, the FORCE simulator did not model countermeasures, thus a participant’s primary recourse when fired upon was to move the asset out of vulnerable range. Sometimes this required a series of steps, such as cancelling an assigned task, creating a new task, and assigning the vulnerable asset to the new task in order to begin moving the vehicle away from the threat. Other times, there was not a designed way to quickly redirect a specific asset at all. Some participants again identified workarounds. For instance, they recognized that they could quickly command a sensor stare point to a point out of the threat range that would force the vehicle to move closer to that point and further from the threat. Understandably, if an asset is being fired upon and a MUM-T system operator becomes aware of this, they need to be able to immediately move the asset out of harm’s way without having to establish a new task or force their way through a cumbersome series of parameters and settings.

C. Subject Matter Expert Observer Ratings

Examining the DECIDE data alongside the raw metrics data reveal that the MU scores for the principal DECIDE tasks were demonstrated to relate to the primary outcomes of interest. The DECIDE MU score was calculated by subtracting the probability of the MU of the task being categorized as room for improvement from the probability of the MU of the task being superior. This score transforms the discrete distribution DECIDE provided into a single number reflecting the strength of task performance. Table 3 displays the correlations between MU scores of the principal tasks and a sample of relevant mission performance outcomes across all vendors’ systems.

Table 3. MU DECIDE Correlations

	Firing AAA Detected	Firing AAA Engaged	Targets Detected	Targets Identified	Threats Tracked	Threats Engage	Threats Destroyed
1. Mission Planning MU	0.345*	0.264*	0.468*	0.342*	0.159	0.337*	0.292*
2. Sensor Payload MU	0.165	0.186*	0.499*	0.479*	0.095	0.469*	0.470*
3. Conduct Reconnaissance Operations MU	0.142	0.206*	0.538*	0.540*	0.100	0.497*	0.488*
4. Tactical Reports MU	0.257*	0.288*	0.464*	0.388*	0.116	0.359*	0.344*
5. Track Target MU	0.305*	0.325*	0.560*	0.564*	0.239*	0.506*	0.525*
6. Store Target MU	0.373*	0.359*	0.465*	0.384*	0.230*	0.388*	0.385*
7. Engage Target MU	0.395*	0.430*	0.695*	0.621*	0.277*	0.676*	0.676*
8. Reacting to Contact MU	0.203*	0.265*	0.0896	0.0281	0.174*	0.024	0.021

Note: * = p is less than .05

For some vendor systems, the DECIDE data did not provide confirmed system performance conclusions for a substantial percentage of the principal mission tasks. This could have occurred for a number of different reasons. First, DECIDE may not have received enough data to judge participant performance because the AMC did not perform the task in a particular vignette. Additionally, contradictory DECIDE data also played a substantial role in limiting the number of firm conclusions stemming from DECIDE. These contradictory DECIDE data arose when the SME judged the technology to serve some participants at a Superior level and some participants at a Room for Improvement level.

Nonetheless, an investigation of the granularities of the available data reveal that the MU scores for Task 3 in Table 3 had a strong positive correlation with detecting targets of 0.538 across all vendors. This indicated that the Task 3 DECIDE scores in Table 3 are related to success in detecting targets. Additionally, comparing MU DECIDE scores for Task 5 in Table 3 with the raw number of threats tracked also resulted in a moderately positive correlation of 0.239. Finally, the MU DECIDE scores for Task 7 in Table 3 are strongly correlated with the raw number of firing Anti-Aircraft Artillery (AAA) engaged at 0.430, targets detected at 0.695, targets identified at 0.621, total threats engaged at 0.676, and total threats destroyed at 0.676. This indicated that the DECIDE scores for Task 7 in Table 3 are related to success in each of these metrics. These findings are not surprising but are nonetheless helpful in confirming DECIDE data and task metric validity.

With the validity of the SME Observer DECIDE MU ratings supported in this way, examining individual vendor system data revealed strengths and weaknesses of each system that were not evident in the objective performance metrics alone. The boxplots shown in Figures 6 and 7 depict the MU scores by subtask for two vendor evaluations and have boxes stretching along the vertical axis from the 25th to 75th percentile of the scores. The whiskers on the plots stretch from the 25th to 75th percentile to the 25th to 75th percentile plus or minus 1.5 times the interquartile range. These whiskers will also stop prematurely if they reach the smallest and largest value in the data.

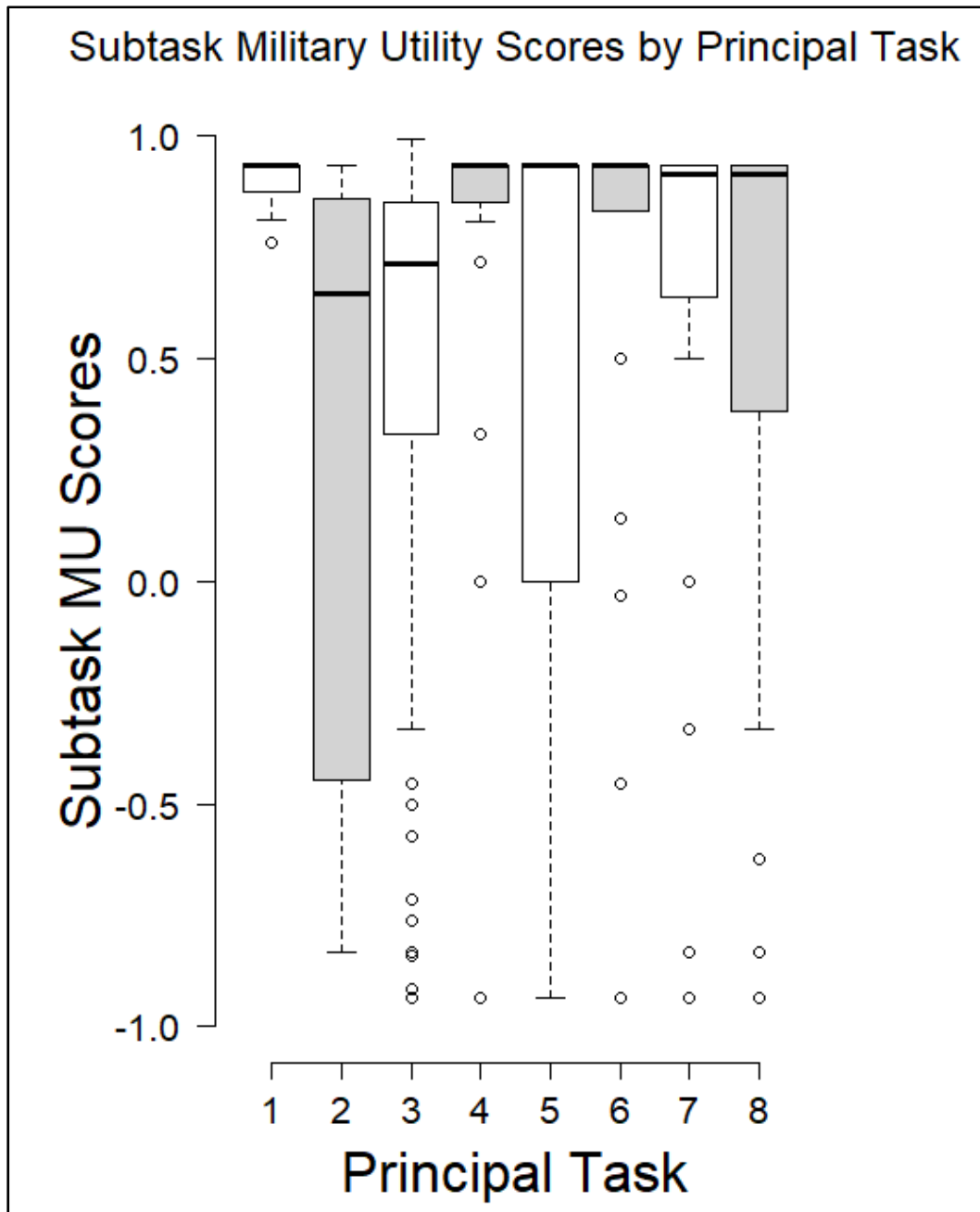


Figure 6. MU Scores by Principal Task, System A

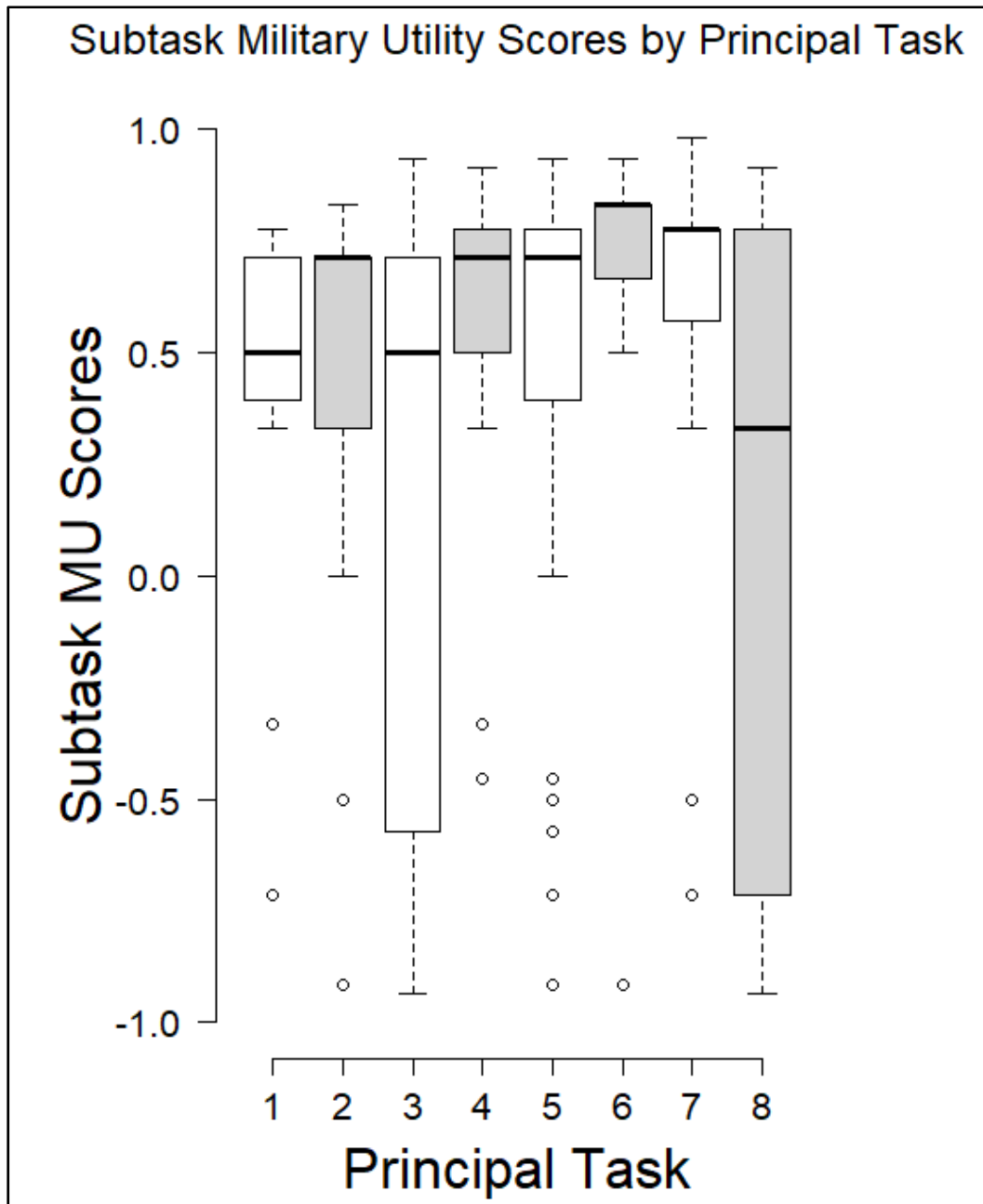


Figure 7. MU Scores by Principal Task, System B

Examination of the granularities of the DECIDE data allowed SUMIT to produce more specific feedback about which technological areas consistently lived up to expectations and which areas were inconsistent. Figure 6 shows, in general, the MU subtask scores reflected positively on this system. The medians of all principal task groups were very positive. While these medians indicate strength in each principal area, the wide boxes and long whiskers present for principal Tasks 2 and 5 indicate that the system did not always excel in terms of operating the sensor payload and tracking targets.

Figure 7 depicts the MU scores for a different system evaluation, which supports the idea that for this system, Tasks 3 and 8 in Table 3 experienced considerably more variability and had more negative subtask scores. Thus, these tasks are potential priorities for improvement for this system due to the inconsistent scores present.

A brief examination of the DECIDE scores split by participant provided some degree of insight into individual participant performance. For example, one system evaluation revealed that one of the participants received mostly negative DECIDE scores, while all other participants had a positive bend to their DECIDE data. These types of results indicated that some participants struggled greatly and were unable to efficiently utilize the system compared to the other participants.

Across all vendors, principal task Requirements Satisfaction (RS) scores were not generally correlated with system capabilities. One pattern that did emerge was the Store Target RS score correlated with task sharing (0.617), ability to react to threats (0.536), and cooperative engagement (0.454). This may be interpreted to suggest that the ability to store targets within the system contributed to successful task sharing, reacting to threats, and performing cooperative engagements. This would seem to make natural sense; however, this post hoc analysis should not be taken as support for the generalizability of this conclusion.

Regarding usability, DECIDE results generally showed moderate to strong correlations between the principal task usability scores and the participant self-report usability questionnaire results, with the exception of Tasks 4 and 8 in Table 3. This general agreement suggests that both the DECIDE data from the SME Observer and the participant responses to the usability questionnaire tended to trend in the same direction, giving evidence that these measures are consistent.

Finally, although some individual systems did show a moderate correlation between DECIDE scores and workload and DECIDE scores and SART SA components, these results did not indicate strong relationships with workload or SA when aggregated across all vendor systems.

D. Situation Awareness

SA as measured by the participant through the self-report SART questionnaire was compared to the same questionnaire completed for the participant by the SME Observer. The data showed mild to moderate agreement between the SME and the participant on the SART Demand characteristics, which were Instability: $r = 0.201$, Complexity: $r = 0.272$, Variability: $r = 0.121$; $p < .05$, but not the Supply or Understanding characteristics or overall calculated SA. This suggests similar opinions between the participants and the SME Observer on the information presented to the participant but disagreement on their ability to process the information and the quality and quantity of the information presented.

Accuracy and mean response latency for the SPAM queries were computed. Participants took a little over 6 seconds to respond to the queries on average, and the responses to the queries were overwhelmingly correct. The response times were not significantly different for any of the three levels of SA queried.

Both the SART and SPAM techniques failed to reflect a complete picture of SA. There were many evaluator-observed occasions in which a participant missed a threat or high value target passing through a sensor field of view while they were looking elsewhere, missed a chat message indicating an asset was taking fire, or missed a new threat icon appearing on the map. These observations not only emphasize the importance of aided target detection and distinct salient alerts for hostile fire and emergent threats but also reflect frequent AMC attentional narrowing [11]. As a result, new threats should be clearly indicated to the AMC as a high priority cue that is certain to be noticed and attended to as necessary, regardless of which information display the emergent threat appears in.

An examination of the agreement between the post-vignette SART questionnaire assessment and the query-based SPAM assessment provided during mission execution revealed very little correlation, further illustrating the difference between a self-report SA rating after the fact and the ability to access specific information in these environments and interfaces to respond to SA queries during mission execution.

Furthermore, there were no identified correlations of significance between SART scores and the primary mission performance metrics across all system evaluations, although there were correlations for individual system evaluations. In one system evaluation, the SART Demand characteristics of instability, complexity and variability were related to higher numbers of threats tracked, engaged and destroyed. On the other hand, Spare Mental Capacity, which is a SART Supply characteristic, and the overall SART score were negatively correlated with these same metrics. Thus, participants who engaged more threats with the system in question likely received lower final SART scores as they rated the scenario as more demanding and noted that they had less spare awareness because they were dealing with more hostiles than other AMCs who engaged fewer targets.

There was also a moderate negative correlation between SART score and the self-reported Performance component of the NASA TLX workload scale ($r = -0.243, p < .05$) across all vendor systems. Note that on the NASA TLX Performance component, higher ratings indicated worse performance. As such, the negative correlation indicated that as performance ratings rose or worsened, the SART levels fell or worsened. This indicated that in participants' own opinions, as their SA improved, their performance improved as well.

Examining the relationship between SA and performance metrics yielded inconsistent results. While this does make broad system agnostic conclusions impossible with the data available, there are still observations of note. Within one system evaluation for instance, the total SART score was negatively related to the number of targets detected by participants. While this correlation was not strong, it did indicate a counterintuitive finding that participants who detected more targets self-evaluated as less situationally aware. An investigation of this relationship revealed that detecting more targets was also positively correlated with the complexity subfactor of the SART. In turn, higher complexity scores lowered overall SART

scores. AMCs who encountered more targets rated the situation as more complex and thus self-evaluated as less situationally aware. This relationship between overall SART scores and targets detected can also be explained through the lens of the workload ratings.

E. Workload

Participant workload in SUMIT was evaluated in three ways:

- Bedford Workload Scale
- NASA TLX questionnaires
- HRV in Round 2

The Bedford workload ratings are a single number representing the workload for the overall primary task of scanning and clearing the ingress route and LZ. The data showed across all system evaluations a moderate correlation between the Bedford rating and the fraction of the LZ scanned ($r = 0.201, p < .05$). No other Bedford correlations of note with mission performance measures were apparent with the data available.

The NASA TLX questionnaire consists of six subcomponents, each rated by the participant on a 21 point scale. The six components are mental demand, physical demand, temporal, performance, effort, and frustration. Typically, these components are weighted in order to compute a total workload score; however, due to the exploratory nature of the SUMIT program, the components were examined individually rather than combined into a composite score. A Principal Components Analysis (PCA) of the workload factors, also including the Bedford ratings, was performed to examine the relationship between these variables, as shown in Figure 8. Results showed that temporal demand, effort, and mental demand were closely related. Interestingly, the Bedford workload rating was closely aligned with the NASA TLX frustration variable.

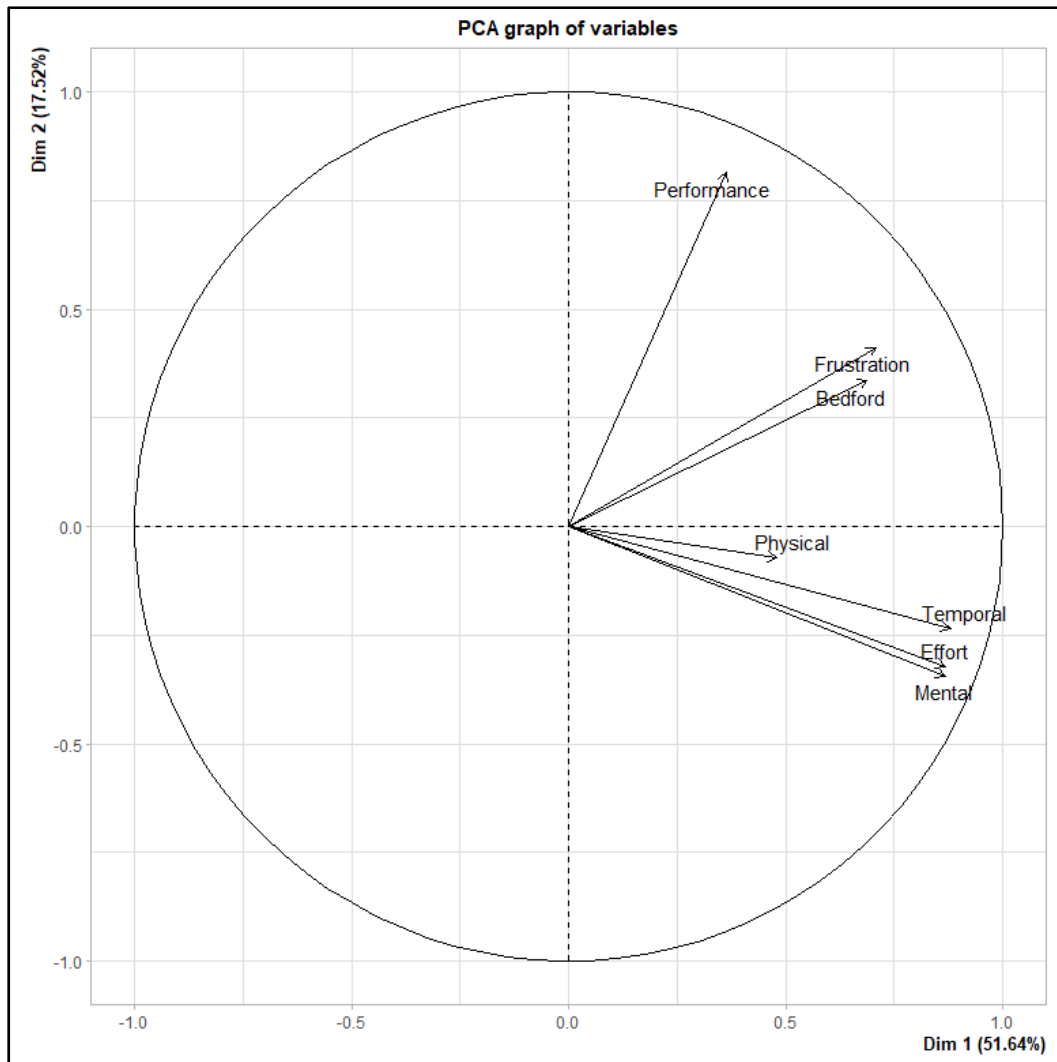


Figure 8. PCA Examining Workload Factors

Examining the Bedford and NASA TLX ratings across vignettes did show variability within select system evaluations. For example, the self-reported workload data captured in Table 4 indicates that the demands placed upon the participants appeared to increase in Vignettes 3 and 4 of Round 2 for this vendor system. This is shown through the increase in the mental and temporal demand variables. This could be partially driven by the fact that the AMCs were in command of more assets during Vignettes 3 and 4 in Round 2. Alternatively, this increase may suggest that this system was particularly ill-suited to help the AMC manage the additional assets.

Table 4. Workload Self-Report Ratings for a Round 2 Vendor System

Vignette	TLX Mental	TLX Physical	TLX Temporal	TLX Performance	TLX Effort	TLX Frustration	TLX Average	Bedford
1	13.25	6.13	11.88	11.75	12.63	13.88	11.58	4.63
2	13.88	7.38	12.38	11.63	13.38	13.25	11.98	4.88
3	15.25	8.25	14.25	11.00	14.50	12.13	12.56	5.25
4	15.50	8.07	13.50	10.50	13.50	12.79	12.31	5.29

Additionally, NASA TLX workload factors, such as temporal and mental demand, appeared to be positively correlated with identifying and engaging more targets for some vendor systems. For systems depicting these results, these positive associations with the metrics data indicated that the participants who found the workload to be more strenuous may actually have been doing more work than their compatriots.

HRV was incorporated into the system evaluations prior to Round 2 as an additional workload measure. Baseline HRV for each participant was collected while they were seated in a small, quiet room planning their mission using paper copies of the mission and mission graphics. An examination of the HRV data yielded no correlation to either the Bedford ratings or the NASA TLX components. Furthermore, workload as measured by HRV was not significantly different between the vignettes. However, extending this comparison to include the baseline data suggests that workload during the baselining period was less than workload during the operational vignettes.

F. Interactions Between Situation Awareness and Workload

One of the drawbacks of the SART is the potential influence of workload. When participants self-reported higher levels of workload, they also self-reported worse SART scores. This appears to largely be driven by the relationship between the workload levels and the SART Demand subfactors. When participants experienced greater levels of workload, they also reported higher levels of demand upon their situational awareness. However, reporting increased demand results in a lower overall situational awareness score using the SART. Thus, when participants reported higher levels of workload, they also generally reported lower levels of situational awareness. Across the consolidated data of all vendors and evaluation rounds, most of the workload variables showed very positive correlations with the SART Demand subfactors, and thus most of the workload variables also have negative correlations with the overall SART score. Tables 5 through 7 show the repeated measures correlations between self-reported workload and self-reported SA across all participants. Stronger correlations ($p < .05$) in Tables 5 through 7 are bolded. These tables demonstrate that the relationships between the NASA TLX workload and SART subfactors were quite strong. This is an important relationship to consider in future work since the level of workload the participants encountered could impact their perceived level of situational awareness.

Table 5. Workload Correlations With SART Demand Factors

	Instability	Complexity	Variability	Demand Total
Bedford	0.039	0.112	0.122	0.099
Mental	0.534	0.635	0.559	0.633
Physical	0.190	0.238	0.260	0.252
Temporal	0.613	0.664	0.562	0.675
Performance	-0.031	-0.064	0.029	-0.024
Effort	0.473	0.595	0.476	0.566
Frustration	0.264	0.214	0.163	0.236

Table 6. Workload Correlations With SART Supply Factors

	Arousal	Concentration	Division of Attention	Spare Mental Capacity	Supply Total
Bedford	0.092	0.199	0.086	-0.505	-0.198
Mental	0.312	0.440	0.278	-0.537	0.030
Physical	-0.184	0.041	0.168	-0.373	-0.250
Temporal	0.156	0.323	0.323	-0.487	-0.018
Performance	-0.173	-0.063	-0.076	-0.136	-0.208
Effort	0.167	0.333	0.240	-0.588	-0.108
Frustration	-0.055	0.070	0.146	-0.282	-0.125

Table 7. Workload Correlations With SART Understanding Factors and Overall SART Score

	Information Quantity	Familiarity	Information Quality	Understanding Total	SART Total
Bedford	-0.203	-0.216	-0.221	-0.261	-0.337
Mental	0.204	0.094	0.121	0.167	-0.333
Physical	-0.082	-0.118	-0.008	-0.082	-0.364
Temporal	0.157	0.041	0.064	0.102	-0.428
Performance	-0.145	-0.228	-0.215	-0.242	-0.243
Effort	0.075	-0.027	-0.001	0.015	-0.452
Frustration	-0.019	-0.226	-0.309	-0.236	-0.381

G. Automation Interventions

One aspect of system autonomy assessed quantitatively was the number of interventions. This was defined as the number of times a participant interrupted the automated behaviors of assigned assets. The data for one system evaluation are shown in Table 8.

Table 8. Selection of Performance Metrics for One System Evaluation Across All Vignettes

Participant	Number of Targets Detected	Number of Targets Identified	Number of Targets Identified as Threats	Number of Threats Tracked	Number of Threat Engagements	Number of Interventions
1	30.75	0.5	0.5	0.25	0.75	7.75
2	70	4.5	4.25	4.25	4.75	14.25
3	42	4.5	3.25	3.25	2	17.5
4	46.33	3.67	1.67	1.67	1	9
5	76.5	5.5	5	5	3.5	14.25
6	53.5	1.75	1.25	1.25	0.75	21.25
7	35.75	0	0	0	0	12.75
8	50.25	4.25	3.25	3.25	3.5	9.5

This metric was not able to be computed for all vendor design solutions since some designs incorporated automation that was always enabled. Nevertheless, these automation interventions helped to paint a more complete picture of the participants' need to modify the asset tasking. Participants able to accomplish their primary tasks with minimal interruptions of the automated system may indicate higher satisfaction with the system generated plans and priorities, reduced need to directly interact with the system, and a willingness to let the automation proceed with the specified tasks. Although there were no statistically significant correlations between task performance and the number of interventions with any one vendor system evaluation, a more holistic data interpretation, including evaluator observations of high performers, is further addressed in Section IV.

H. Trust and Usability

The Schaefer Trust in Automation questionnaire, as shown in Appendix E, was administered as a pretest after the participants had received the PowerPoint classroom training but before they had interacted with the system at all. The questionnaire was again administered at the end of the second day, after the fourth and final vignette. Results showed that on average, the participants' trust in automation decreased from pretest to posttest for all of the systems. This is shown in Figure 9, where all of the positively scored questions decreased in response value, while the negatively scored questions increased in response value. This suggests that the technology did not meet the expectations of the participants in some respects, which is likely due to system glitches and partial software crashes often inherent in innovative S&T.

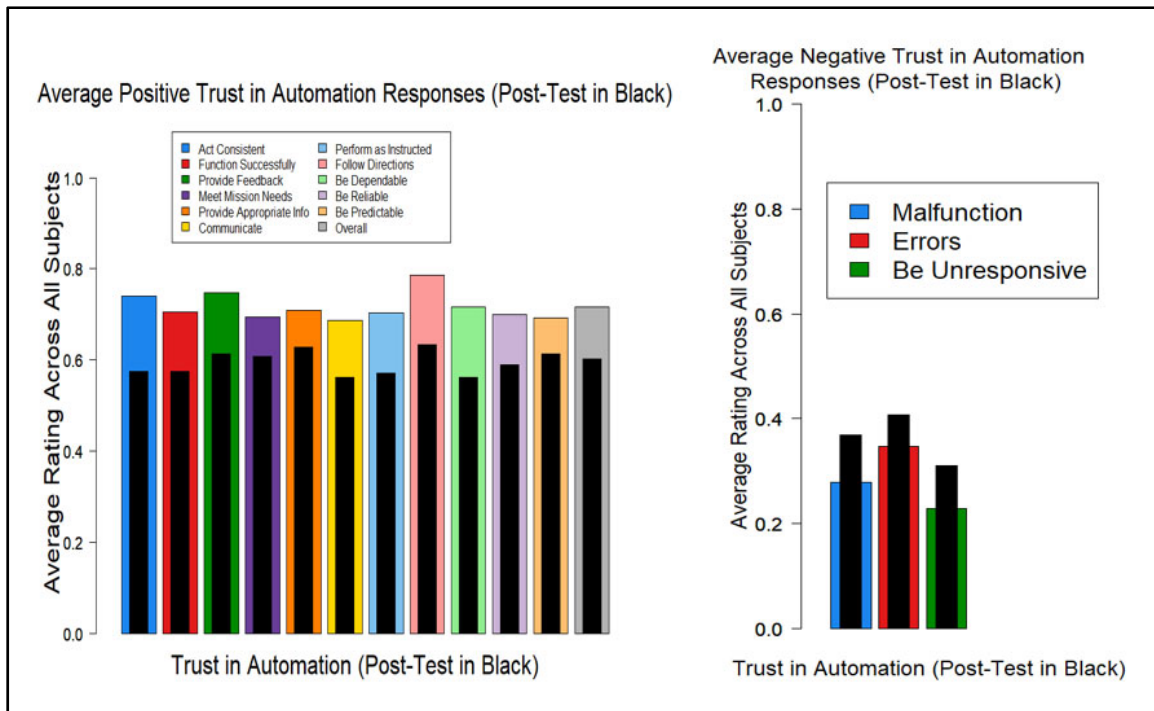


Figure 9. Trust in Automation Responses

Examining the trust in automation at an individual vendor system level provided more detailed insights on the relationships between trust in automation and other variables. The following paragraphs in III.H provide results of selected individual vendor evaluations rather than overall findings.

In some cases, the pretest trust in automation measures showed strong positive ties to performance metrics, such as targets detected, threats engaged, and threats destroyed. Thus, participants who had more positive expectations for the technology before the study began yielded more positive results. These relationships demonstrated that participants who placed more confidence in the technology at the outset of the experiment ended up performing better on average over the course of the four vignettes.

Figure 10 depicts a correlation plot for one vendor system evaluation with posttest trust in automation factors on the left and usability questions along the top. Note that the scale for the usability metrics was flipped for the plot such that better trust in automation scores and better usability scores would positively correlate rather than negatively correlate. AMC trust in automation responses were generally closely related to their usability questionnaire responses. The predominately blue correlation plot demonstrates that the relationships between these two sets of variables are quite strong, and further examination of individual correlations may be desired on future MUM-T system evaluations.

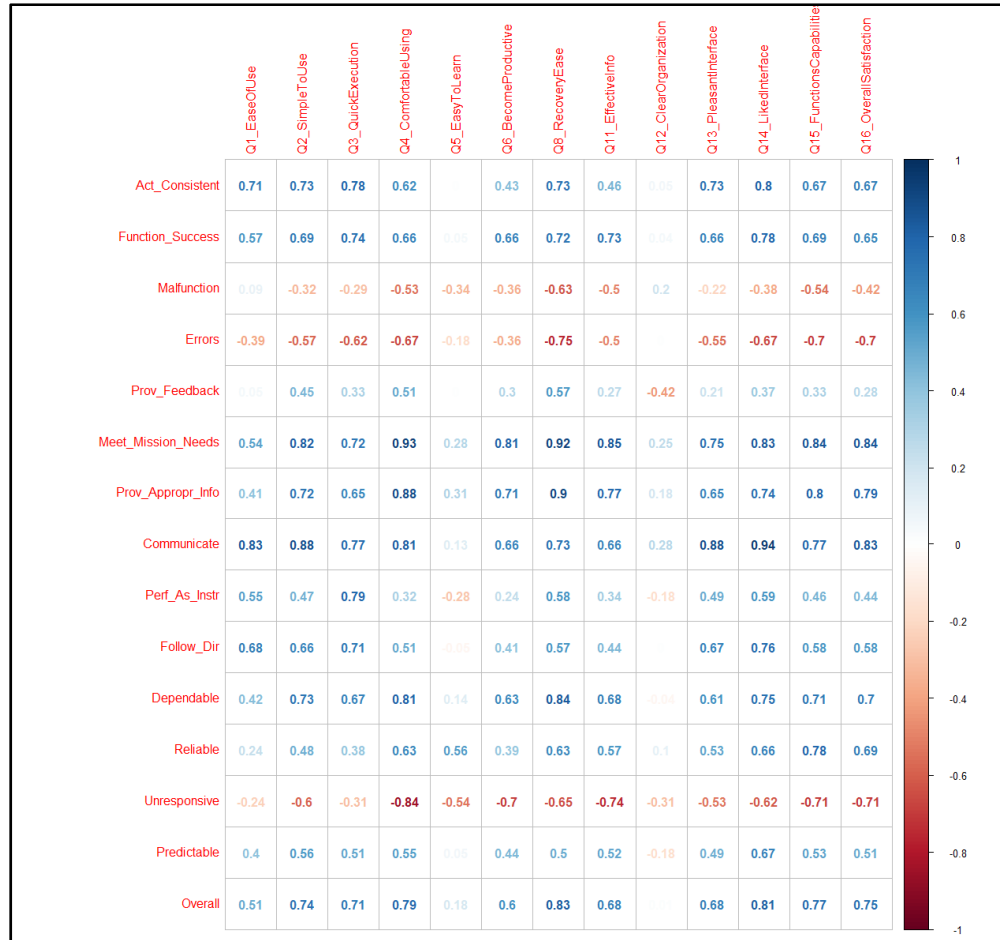


Figure 10. Trust in Automation Posttest and Usability Correlation Plot

For another vendor system evaluation, however, although the trust in automation decreased from pretest to posttest, the posttest usability scores of the participants were still quite positive, indicating that the technology was still considered quite usable regardless of the degraded trust on the participant's behalf. Interestingly, the trust in automation and usability scores for the two weakest performers provide sharp contrasts to each other. While one participant rated the usability of the system and their trust in automation as very low, a second low performer gave much higher scores on usability and trust in automation. Both subjects appeared to perform poorly in the metrics, but they came away with diametrically opposed views on the adequacy of the technology.

Lastly, the DECIDE data indicated the strength of this same system across a multitude of tasks. Even though large differences in average scores exist between participants, the participant with the lowest MU and RS scores still had scores that reflected very positively on the system. These consistently positive scores across all participants for this system indicated that the technology was very effective in the DECIDE task areas. All of the firm conclusions in the DECIDE data for this system evaluation identified the system as supporting task performance at a superior level. While the DECIDE data indicated the strength of the system across a multitude of tasks, the performance metrics did not support the success of the technology as strongly.

Overall, very few of the metrics met the benchmarks that the SMEs had established as expectations for participant performance to be rated as Sufficient and Superior for each vignette. This could be due to many different reasons. For example, SME expectations were unrealistic, participants underperformed, technologies did not meet SME expectations, metrics were not measured accurately, and so forth.

I. Age

Examining the relationship between age and other variables revealed several insights. Although again there were no significant correlations across all evaluations, within the individual vendor system evaluations, age repeatedly appeared to be an influence.

One example of this influence extracted from an individual vendor system evaluation can be seen in the number of targets in each step of the target resolution process when split by age, as shown in Table 9. For this system evaluation, two participants were over the age of 50 while six were under the age of 40. Note that in this sample, age was also strongly correlated with number of flight hours ($r = 0.89$), and the two oldest AMCs in this sample were also the only two non-active duty Soldiers. Thus, these potentially impactful variables are impossible to disentangle from age for this limited dataset.

Table 9. Target Resolution Metrics for One Evaluation by Age

Metric	Older AMCs	Younger AMCs
Targets Detected	3.25	4.92
Targets Identified	1.75	4.00
Targets Tracked	0.88	2.96
Targets Engaged	0.00	2.00
Targets Destroyed	0.00	1.83

Examining the DECIDE data alongside of the raw metrics data revealed that, for this system evaluation, the MU scores for the principal DECIDE tasks were correlated with targets detected, tracked, and destroyed. Thus, DECIDE scores were demonstrated to relate to these outcomes of interest. Figure 11 displays the task scores while examining the role of age in the DECIDE data. The tasks in this plot are split by age with AMCs less than 40 years of age appearing in the white boxplots and AMCs greater than 50 years of age appearing in the gray boxplots.

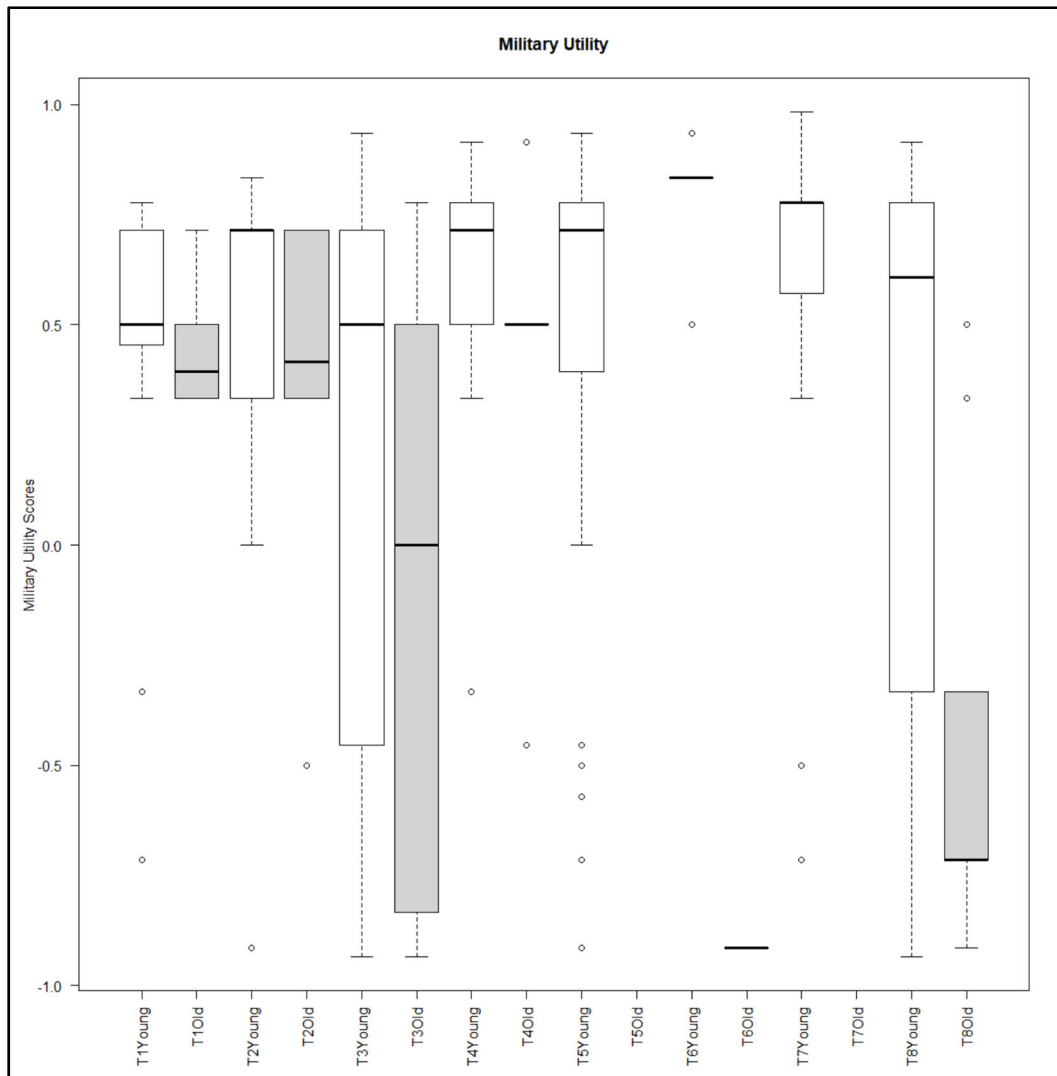


Figure 11. DECIDE MU Scores by Task, Split by Age for One System

Figure 11 demonstrates that the younger AMCs consistently had higher MU DECIDE scores. The median scores for the younger AMCs in each task are higher than the median scores for the older AMCs for the same tasks. Thus, the younger AMCs consistently produced higher MU scores across all principal task categories.

In a separate vendor system evaluation example, age of the participants was relatively closely linked to the workload ratings as measured by NASA TLX. The two oldest participants in this evaluation had the two highest average levels of effort, temporal demand, and physical demand across the four vignettes. They also both had higher than average levels of mental demand and frustration. These high averages across most workload metrics illustrate that the older AMCs rated the workload as more strenuous than their younger counterparts.

Age also appeared to have relationships with the situational awareness data as measured by SART. The two oldest participants in the experiment rated the situational awareness Demand characteristics higher and noted that they had less spare mental capacity than the other AMCs on average. Overall, the relationships between age, workload, and situational awareness indicated that the older participants generally found the experience to be more demanding than the younger participants.

While older participants rated the workload as more strenuous and appeared less situationally aware than their younger colleagues for this system, the DECIDE scores of the younger and older AMCs were largely similar. Note, however, that for this particular system, the SME Observer DECIDE ratings indicated that the technology was remarkably supportive for the user to accomplish the tasks across all participants and task areas. As there is very little variability in the overwhelmingly positive DECIDE scores, it is difficult to assess potential ways that AMC demographics, such as age, might interact with the task performance.

IV. DISCUSSION

A. Holistic Data Interpretation

Due to the enhancements made to the FORCE simulators and vendor system design solutions between evaluations, periodic partial system crashes, and small sample size of eight participants per evaluation, the quantitative results from the individual SUMIT evaluations cannot be generalized across all MUM-T systems. Nonetheless, when examined from a holistic perspective considering the quantitative data, SME ratings, evaluator observations, and participant feedback together, many informative suggestions for future MUM-T system evaluations can be extracted.

The number of times that the UAS assets were targeted, fired upon, and shot down increased as the vignettes progressed. These figures are consistent with the increasing number of assets that were available to the AMC and emphasize the importance of SA during mission execution. Providing an operator with additional assets without the appropriate conveyance of SA information, such as incoming fire, will result in less than efficient use of the assets and negate much of the potential return on investment.

On average, trust in automation decreased from pretest to posttest for all of the systems. This primarily depicts the presence of software issues and partial system crashes disrupting the operator trust; however, participants who rated their pretest trust in automation as higher ended up performing better on average over the course of the four vignettes. Participants who had more positive expectations for the technology before the study began ended up having more positive results.

Posttest trust in automation and usability were generally, but not always, correlated. Examining the relationships in situations where this was not the case indicated participants found systems with fewer required interactions to be more usable. This did not necessarily lead to increased performance. Although the participants in this evaluation were given several hours of hands-on training in the simulator until demonstrating a level of comfort and proficiency to begin the training vignettes, these were still systems they were learning over time. More time with the

system could have enabled more complex systems to be seen as more usable. With caution that simpler does not mean fewer capabilities, reduced perceived system complexity was evident in participants working with systems requiring fewer interactions, but more complexity per the SART subfactor was correlated with finding more targets and engaging more threats as well as higher workload and higher performance. It is likely participants perceived themselves as being in more variable and complex situations because they detected more targets and thus had to neutralize additional threats. AMCs who rated the situations as less variable and less complex were likely missing some of the complexities of the situation through failure to detect as many targets as their peers.

Further examination of the relationship between SA, workload, and performance metrics revealed that participants who had rated the workload as higher and their SA as lower did generally have better performance metrics than other participants. This indicates that these participants did actually experience a greater workload than the other participants and thus their lower SART scores may simply be due to the fact that they had to perform more tasks and so they had less spare mental capacity and experienced more instability, complexity, and variability in the situations that they were engaged in, resulting in lower overall SA scores as assessed by the SART.

LOA was not correlated with mission performance. A highly automated system may result in fewer interactions with the system and subsequently higher usability ratings, yet the systems that could not provide the ability for a user to interrupt the automation and quickly move an asset out of harm's way or manually direct a sensor were not perceived to be sufficient. LOA is not a recommended metric for MUM-T system assessment.

The threat engagement-to-intervention ratio was particularly interesting when analyzed together with SME and evaluator observations. Note that interventions are different than system interactions. Any input to guide a sensor for instance could be considered a system interaction. Interventions, on the other hand, were defined as interruptions of the automated behaviors of assigned assets. Participants considered to be high performers based on SME and evaluator observation also had the highest threat engagement-to-intervention ratios. In other words, they were more efficient in tasking the vehicles, spending less time locating and classifying targets and retasking vehicles, and more time engaging targets identified as threats. This threat engagement-to-intervention ratio has emerged as a candidate for further inclusion and examination in future MUM-T system evaluations.

One primary takeaway not reflected in any quantitative measure was the observed differences between participants using the autonomy versus fighting against the autonomy. Some systems had very high levels of autonomy such that the designers expected the AMC to always let the system decide which areas to search and develop a flight and sensor plan for searching. Many participants were uncomfortable with this and continuously tried to specify which system would perform which task, even going so far as gaming the system with workarounds in order to force this functionality onto a system that did not natively support it. Whether the system reached the same decision on asset availability and task prioritization seemed to be irrelevant since these participants did not give the system a chance to make these determinations. This desire for direct asset management and control appeared to be rooted in experienced decision making strategies based on considerations not modeled in the simulation,

such as asset detectability due to noise. In theory, this high LOA lifted the detailed control responsibility from the participant, but in practice, this struggle against the automation led to tremendous frustration on the part of the participants as well as the system designers.

One of the demographics questions asked about recent video game experience and types of video games played. Although there were no statistically significant results generated from this question, note that of the 39 participants, only one self-identified as having recent experience with Real-Time Strategy (RTS) video games, and this participant was an extremely quick learner and high performer on their system evaluation, all the while commenting on how similar this particular technology was to RTS games, such as StarCraft and Command & Conquer, and how the participant felt this helped to understand the system functionality. They were able to easily multitask while rapidly interacting with the touchscreen technology, developed novel strategies to maximize the use of the system, efficiently tasked the unmanned assets, and often reached accurate conclusions about how the system functioned even before being trained on the features.

The differences in performance due to age should be interpreted in the context of compressed training time. Evaluators and trainers alike observed that older participants struggled to retain all of the training information regarding a system's features and functionalities. The older participants typically appeared to have a shallower level of understanding regarding a system's autonomous systems, capabilities, and context dependent behaviors as well as trouble recollecting the operating procedures for the human machine interface in general. It appeared that with the compressed training time, many of the older participants either could not retain the abundant amount of training material or were unable to apply this information in the task saturated and temporally demanding simulation environment. That is, critical parts of the training seemed to be discarded once the vignette started. Operationally, fielded MUM-T systems would naturally have extensive and rigorous training sessions to avoid these types of consequences, perhaps minimizing differences in performance due to age.

The strong correlation between age and flight hours ($r = 0.81$ across all participants) was noted, and the specific effect of increased flight hours was occasionally very apparent. One particularly experienced participant repeatedly tried to engage targets using trained motor pathways and had great difficulty reframing their mindset away from experience and muscle memory, including one occasion staring at a target trying to remember how to engage for several minutes. The participant seemed unaware at times that he or she was not in a traditional aircraft, trying multiple times to perform tasks, such as target engagement, in an expected way without realizing a novel and innovative design concept was being used. This may raise a question as to how much new training might be required of seasoned pilots who are adapting to future MUM-T systems.

B. Design Feedback

Though the SUMIT program was not specifically focused on usability enhancements of the vendor systems, many suggestions and observations based on system design impacted user decision making, SA, and performance. For example, Round 2 evaluations provided multiple weapon types to choose from, such as Hellfire-R missiles, APKWS rockets, and LRPF. The

icons for the missiles and rockets in one system appeared similar in design and were grouped closely together, resulting in participants selecting ineffective weapons for target engagement. Participants knew that engaging an armored tank with an APKWS rocket was ineffective, and this was accurately represented in the simulation as having no effect, yet after seeing no effect and without realizing their error in weapon selection, the participants would suspect a software glitch or other error and attempt the same incorrect process again. Salient differences between weapons of different capability would be very helpful to avoid this situation and increase participant SA.

An example of a design element found to be beneficial is the scribble tool implemented in one vendor's system design. As the HMI in these systems consisted of touchscreens, the quickest and easiest way to add an irregularly shaped area to the map was using a finger to draw the area on the map. Participants used this feature constantly to draw new search areas based on incoming information, draw a no-fly zone around an identified threat, or split a large area into smaller prioritized search regions. The improved efficiency of drawing directly on the map as opposed to specifying parameters through a menu system was dramatic and instantly recognized as desirable and beneficial by the participants.

An important design element in many complex interfaces is the ability to filter or declutter information from a busy screen. When multiple icons or graphical elements are in close proximity, it can be difficult to select or even see the desired icon amongst the clutter. One implemented declutter design was very capable and customizable—so much so that it became cumbersome. Providing a large number of filter options while spending a relatively short amount of time training participants on all of the options had three negative consequences. First, participants would forget how to access the declutter tool. Second, once they accessed the tool, they would click through many icons or combinations of icons via trial and error in order to hide the specific icons or clutter they intended to hide. Lastly, participants would forget that they had turned on a data filter and thus not have that information available to them at a later time when it was beneficial or necessary.

The impact of an individual decision aid can be difficult to quantify, yet observations and participant feedback identified many information elements that indeed aided participant decision making. Providing a mission status panel that showed all tasks currently underway and which assets were performing them was very informative and received positive remarks from participants, particularly in comparison to the same system depicting each asset and their assigned task. The focus on the task perspective, rather than the asset perspective, was the primary takeaway. In a complex multitasking environment with automation controlling multiple assets, the ability to see in one location, the status of all tasks, and the assets assigned to them helped aid participant decisions on task assignment, prioritization, and adaptation.

Another impactful decision aid requested by the participants was a clarification between effective and ineffective fire on the teamed assets. When removed from direct perception, an operator relies on visual or auditory information displays to identify when an asset is being fired upon. Participants responded to notifications of assets under fire differently, depending on the specific asset type and current mission priorities. Clarifying effective versus ineffective fire can provide additional information to aid in this decision. Just because an asset reported it was being fired upon did not always demand an action or response from the

participant, and this additional information would increase SA and further inform decision making.

The importance of alerting the operator to new targets cannot be overstated. Since there is no reliable prediction of where on an interface an AMC might be looking at a given point in time, it is important that the alerts representing a new threat are salient and attention getting. Often this piece of information will influence an AMC's priorities and decision making instigating modifications to their mission plan depending on the threat level and location.

Other relevant design elements requested by the participants included the ability to drop a pin on the map to investigate later, providing a zoom-based sensor slew rate, an indication for loss of asset signal, and the ability to digitally transmit a grid coordinate of a position on the map, such as a sensor stare point, without manual data entry. For example, copy and paste or click could be used to send an MGRS location from the chat window without manually typing the coordinate.

Any situation in which the automation violated or conflicted with the participants' experiences and decision making process negatively impacted their performance to some degree. A common example of this was exhibited when directing a sensor to stare at a location. Participants consistently expected the UAS platform to fly near the spot but stay at an offset distance. Ultimately, some system designs guided the asset to the spot to orbit directly over it. This negatively impacted the visibility of the object by providing a top-down perspective instead of a slanted viewing angle and unnecessarily increased asset detectability and vulnerability by moving them directly overhead of a potential threat.

C. Air-Launched Effects

Participants were provided with four ALE platforms in the Round 2 evaluations. Participants were instructed on how to launch the ALEs and informed of their capabilities, but they were intentionally not provided any guidance or suggestions on the best way to utilize them. Participants responded by incorporating the ALE platforms into their mission plan without any special treatment. Essentially, they were just another asset at their disposal, indistinguishable from other teamed UAS assets. The platform capabilities became the primary considerations for asset employment. Capability factors that informed AMC decision making included payload, position, speed, detectability, and maneuverability.

Software constraints had varying effects on participant employment of ALEs. For instance, in one system design, the ALE upon launch would immediately begin descending or ascending to a predetermined altitude in a tight orbit pattern. Once reaching the nominal altitude, only then would the ALE begin moving toward the commanded location. This design negated a primary anticipated benefit of the ALE platform, which was to remain stowed upon the ownship or larger UAS mothership until needed and then quickly be available for tasking. Participants using this system quickly recognized that the ALE platform would take several minutes before beginning movement toward their commanded location and considered this amount of time unacceptable, thus they adapted to this system design by immediately launching the ALEs upon mission start.

The ALEs within the SUMIT evaluation did not carry a laser designator onboard and were thus not able to be used for cooperative target engagement. Rather, ALEs were utilized for two primary purposes: performing reconnaissance and drawing enemy fire to reveal threat locations.

Once an ALE discovered an entity and the aided target recognition informed the participant, the participant would attempt to further classify the entity. If the entity was neutral, it would be ignored and scanning resumed. If the entity was deemed to be a moving threat, the participant would keep the camera fixated on the target as much as possible until another asset with additional capability could arrive to designate for cooperative engagement or engage the target. If the entity was stationary, the ALE could observe while an LRPF mission was conducted to eliminate the threat and subsequently conduct battle damage assessment. Overall, the ALEs were well-received and perceived as very useful by the participants as another sensor through which to gain additional situation and battlespace awareness. They were considered for utilization in the same way that more traditional UAS assets were and sometimes forgotten about or not launched at all.

D. Degraded Communications

The SUMIT program imposed additional constraints on the vendors in Round 2 to begin to explore a degraded communications environment. Communication loss between the ownship and UAS assets in an operational environment can be intentional, such as reduced asset signal detection, or unintentional, such as electronic warfare, jamming, or LOS interruptions. Regardless of the initiating reason, when considering the future fight against near-peer threats, a MUM-T system will need to abandon the technological assumption of multiple high bandwidth video feeds streaming simultaneously into a cockpit. The specific requirement change in the SUMIT Round 2 evaluations meant that the HMI was allowed to present to the participant the ownship sensor and a maximum of one other sensor video at a time. Vendor designs accommodated this constraint in one of two ways:

- Only presenting two sensor feeds on screen that the participant could switch out as desired
- Still showing all sensor feeds simultaneously but only showing live video for the ownship and one other asset, while the remaining sensor feeds were periodically updated still photos

A total of three vendor design solutions met this constraint—one Round 1 vendor and both Round 2 vendors—giving the evaluators ample opportunity to identify strengths and weakness of these approaches. The primary weakness of the first approach was a loss of SA. If an entity was detected by an asset not displaying its sensor on screen, the participant relied on notifications from the system to switch to that sensor feed and identify the entity. Whether a drag and drop approach or a quick tap of the touchscreen seemed essentially irrelevant as long as the participant was able to quickly and easily switch to their desired sensor view. This design also necessitated a reliable entity detection notification or cue for the sensors not currently displayed on screen to inform the participant of detected entities that they would not otherwise have seen.

With the second approach, which showed live video for two sensors and periodically updated photos of the last image seen in the other payload views, the potential gains in SA appeared to be lost by consistent participant modal confusion. Participants were often unable to discern whether the sensor was displaying a still photo or live video fixated on a stationary object. This confusion resulted in degraded performance, unnecessarily repeated inputs, and participant frustration when trying to interact with the sensor feeds. If live videos are presented in combination with still images or previously recorded videos, the clear identification of which videos are live is critical to maintaining operator awareness and decreasing response time and modal confusion of operators.

V. CONCLUSIONS

The SUMIT evaluations served two primary purposes:

- Identify capabilities for a MUM-T system that would increase combat effectiveness
- Provide recommendations on metrics through which to assess these systems

Utilizing a combination of quantitative data, participant feedback, and SME observation, the evaluators were able to accomplish these goals. The capabilities identified as most impactful in the five SUMIT vendor system evaluations were aided target recognition, cooperative engagement, high-level tasking, task sharing, task queueing, information at the point of need, automation transparency, sensor scan history and current scan status, notification of task completion, immediate direct control availability, and the ability to quickly react to emergent threats.

Regarding relevant metrics, the importance of context and expert opinion cannot be overstated. Improved AMC combat effectiveness is the end goal, but there is no one overall mission performance metric that can be recommended for all missions utilizing MUM-T systems. An acceptable number of targets detected, identified, engaged, and destroyed will vary greatly depending on context and AMC priorities. Failing to accomplish one task may be acceptable if many other tasks are accomplished instead.

To evaluate MUM-T mission performance, future MUM-T system evaluations should first define task priorities. Once the primary task priorities are clearly defined, one recommendation is to rely upon input from SMEs familiar with the mission set to provide relative weighting to secondary tasks. Calculating performance in these individual tasks, such as percentage of relevant area covered for a reconnaissance task, multiplied by the weights assigned by the SMEs will enable overall contextual mission performance as a metric on which to evaluate MUM-T systems.

While system designers often rely on workload as a measure of system effectiveness, an AMC finding and engaging more targets typically equates to performing more tasks, often leading to higher self-reported workload ratings. SUMIT evaluations showed in this context that higher workload was correlated with greater demand on SA, which resulted in lower self-reported SA ratings overall. Thus, a better MUM-T system might have higher workload ratings and lower SA ratings than a worse system.

With that said, for the SUMIT mission vignettes, the tasks most critical to successful mission performance were reconnaissance and target engagement. MUM-T systems regarded as lacking in support for these specific tasks led to difficulty detecting, identifying, locating, and destroying threats and were generally regarded with low usability and low trust in the automation. It is recommended that future MUM-T system designers focus on ensuring that these priority AMC mission tasks can be performed in a usable and capable manner.

LOA is essentially meaningless if an operator cannot also take immediate manual control during critical events. Time to complete reconnaissance is not a recommended metric simply because this duration fluctuates depending on the asset or assets assigned, which in turn varies depending on AMC priorities.

The task to intervention ratio was identified as a metric that may reflect how easily an operator is able to use a MUM-T system. The specific task should be aligned with the SME identified task priorities. For the MUM-T missions performed by an AMC in the future fight, this could be targets engaged per intervention and reconnaissance tasks completed per intervention. Interventions in this context refer to interrupting an automated system or behavior. Taking manual control of a sensor to redirect would be considered an intervention, while fine tuning sensor payload orientation with touchscreen presses would not be considered multiple interventions. This task to intervention ratio metric avoids the issue of relying on a higher raw number of tasks completed to indicate better performance, instead asking how easily an operator can complete the task whenever they deem it a priority to do so.

For multi-vehicle MUM-T systems, the efficiency of asset utilization is key. Whether the vehicle is idly loitering unnoticed while waiting for a task or the sensor is staring at a patch of earth where a target used to be, both are examples of inefficient asset utilization. Asset utilization efficiency for the sensor and platform are recommended metrics as indications of MUM-T system performance.

These evaluation conclusions emphasize the value of Soldier touch points in S&T programs. Their expertise and operational familiarity lent added operational validity to the findings. Participants were eager to provide feedback on capabilities for an AMC using MUM-T systems and were incredibly insightful and forthcoming with their information processing and decision making processes.

This valuable input was especially timely, given that the Army has identified speed and scale as two of the core challenges of the future fight. There cannot be an assumption of the luxury of time for military decision making, and the number of sensors available on the battlefield is only expected to grow. The Warfighter of the future must be able to leverage this information and data more efficiently than their adversaries. The capability and metrics recommendations elicited from their experiences in the SUMIT program and described in this report will inform MUM-T system design and assessment, which will continue to ensure technological overmatch in future multi-domain operations.

REFERENCES

1. “The U.S. Army robotic and autonomous systems strategy,” United States (U.S.) Army Training and Doctrine Command (TRADOC) , 2017, https://www.tradoc.army.mil/portals/14/documents/ras_strategy.pdf
2. Snyder, M. I.; Hall, B. T.; and Alicia, T. J., “Use of the FACE™ technical standards in SUMIT: Lessons learned and best practices from GE Aviation and the U.S. Army in SUMIT,” Paper presented at FACE™ & SOSA™ Expo and Technical Interchange Meeting, Dayton, OH, September 2019.
3. Alicia, T. J. et al., “Removing the bottleneck: utilizing autonomy to manage multiple UAS sensors from inside a cockpit,” Proceedings of SPIE Volume 10640, Unmanned Systems Technology XX, 106400L, doi:10.1117/12.2303915, 03 May 2018.
4. “STANAG 4586,” Edition Number 4, North Atlantic Treaty Organization (NATO) Standardization Agreement 4586, NATO Standardization Office (NSO), 05 April 2017.
5. “Aviation flight regulations,” Army Regulation 95-1, United States (U.S.) Department of the Army, 2018, https://armypubs.army.mil/productmaps/pubform/details.aspx?pub_id=1003624
6. Durso, F. T. et al., “Situation awareness as a predictor of performance for en route air traffic controllers,” Report Number DOT/FAA/AM-99/3, Office of Aviation Medicine, Washington, DC, 1999.
7. Endsley, M. R., “Toward a theory of situation awareness in dynamic systems,” Human Factors, Volume 37, pp. 32-64, doi:10.1518/001872095779049543, 1995.
8. Tao, D. et al., “A systematic review of physiological measures of mental workload,” International Journal of Environmental Research and Public Health, Volume 16, Number 15, p. 2716, doi:10.3390/ijerph16152716, 2019.
9. Shaffer, F. and Ginsberg, J. P., “An Overview of Heart Rate Variability Metrics and Norms,” Frontiers in Public Health, 28 September 2017, <https://www.frontiersin.org/articles/10.3389/fpubh.2017.00258/full>.
10. Chen, J. Y. C. et al., “Situation awareness–based agent transparency,” Report Number ARL-TR-6905, Army Research Laboratory (ARL), Aberdeen Proving Ground, MD, United States (U.S.), 2014.
11. Prinett, J. C.; Mize, A. C.; and Sarter, N., “Triggering and detecting attentional narrowing in controlled environments,” Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Volume 60, Number 1, pp. 298-302, doi:10.1177/1541931213601068, 2016.

LIST OF ABBREVIATIONS, ACRONYMS, AND SYMBOLS

%	percent
<	less than
AAA	Anti-Aircraft Artillery
ALE	Air-Launched Effect
ALTIUS	Air-Launched, Tube-Integrated, Unmanned System
AMC	Air Mission Commander
APKWS	Advanced Precision Kill Weapon System
AvMC	Aviation & Missile Center
BAA	Broad Agency Announcement
COIN	Counter-Insurgency
CPG	Copilot Gunner
DAM	Decentralized Asset Manager
DATE	Decisive Action Training Environment
DECIDE	Diagnostic Engine for Cognitive Improvement and Decision Effectiveness
DEVCOM	Combat Capabilities Development Command
FACE	Future Airborne Capability Environment
FLOT	Forward Line of Own Troops
FORCE	Future Open Rotorcraft Cockpit Environment
FVL	Future Vertical Lift
GE	General Electric
HF	High Frequency
HMI	Human Machine Interface
HRV	Heart Rate Variability
Hz	hertz

LIST OF ABBREVIATIONS, ACRONYMS, AND SYMBOLS (CONTINUED)

km	kilometer
knot	kn
LF	Low Frequency
LOA	Level of Automation
LOI	Level of Interoperability
LOS	Line of Sight
LRPF	Long-Range Precision Fire
LZ	Landing Zone
METT-TC	Mission, Enemy, Terrain, Troops Available, Time, and Civilian
MOMS	Mission Operation Management System
MOSA	Modular Open Systems Architecture
MGRS	Military Grid Reference System
MU	Military Utility
MUM-T	Manned-Unmanned Teaming
N/A	Not Available
NASA	National Aeronautics and Space Administration
NAVSEA	Naval Sea Systems Command
PCA	Principal Components Analysis
PNS	Parasympathetic Nervous System
PSSUQ	Post-Study System Usability Questionnaire
Qual	Quality
Quan	Quantity
recon	Reconnaissance
RS	Requirements Satisfaction

LIST OF ABBREVIATIONS, ACRONYMS, AND SYMBOLS (CONCLUDED)

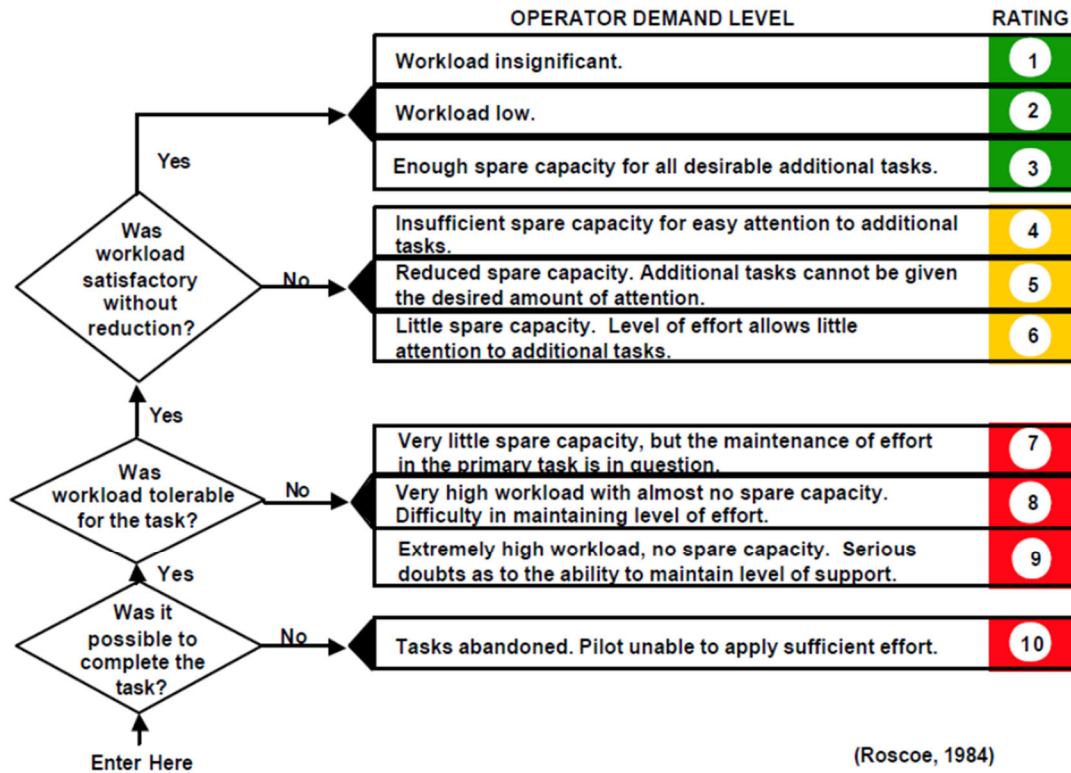
RTS	Real-Time Strategy
RWR	Radar Warning Receiver
S&T	Science and Technology
SA	Situation Awareness
SART	Situation Awareness Rating Technique
SCORCH	Supervisory Controller for Optimal Role Allocation for Cueing of Human Operators
SME	Subject Matter Expert
SMS	Simulation Management Station
SNS	Sympathetic Nervous System
SPAM	Situation Present Assessment Method
STANAG	Standardization Agreement
SUMIT	Synergistic Unmanned Manned Intelligent Teaming
TAI	Target Area of Interest
TDM	Teaming Data Manager
TECUMSA	Tasking and Execution of Collaborative Unmanned and Manned Systems with Autonomy
TLX	Task Load Index
TOC	Tactical Operations Center
U.S.	United States
UAS	Unmanned Aerial System
UAV	Unmanned Aerial Vehicle

APPENDIX A
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION TASK LOAD INDEX

Rating Scale Definitions

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

APPENDIX B
BEDFORD WORKLOAD SCALE



APPENDIX C
SITUATION AWARENESS RATING TECHNIQUE

Rate the level of each component of situation awareness that you had when you performed mission commander tasks during the mission that you just completed. Circle the appropriate number for each component of situation awareness (e.g., complexity of situation).

DEMAND

Instability of situation: Low 1-----2-----3-----4-----5-----6-----7 High

Variability of situation: Low 1-----2-----3-----4-----5-----6-----7 High

Complexity of situation: Low 1-----2-----3-----4-----5-----6-----7 High

SUPPLY

Arousal: Low 1-----2-----3-----4-----5-----6-----7 High

Spare mental capacity: Low 1-----2-----3-----4-----5-----6-----7 High

Concentration: Low 1-----2-----3-----4-----5-----6-----7 High

Division of attention: Low 1-----2-----3-----4-----5-----6-----7 High

UNDERSTANDING

Information quantity: Low 1-----2-----3-----4-----5-----6-----7 High

Information quality: Low 1-----2-----3-----4-----5-----6-----7 High

Familiarity: Low 1-----2-----3-----4-----5-----6-----7 High

SART Definition of Terms

Situation Awareness is defined as “timely knowledge of what is happening as you perform your mission commander’s tasks”

DEMAND	
Instability of Situation	Likelihood of situation to change suddenly
Variability of Situation	Number of variables which require your attention
Complexity of Situation	Degree of complication (number of closely connected parts) of the situation
SUPPLY	
Arousal	Degree to which you are ready for activity
Spare Mental Capacity	Amount of mental ability available to apply to new tasks
Concentration	Degree to which your thoughts are brought to bear on the situation
Division of Attention	Amount of division of your attention in the situation
UNDERSTANDING	
Information Quantity	Amount of knowledge received and understood
Information Quality	Degree of goodness or value of knowledge communicated
Familiarity	Degree of acquaintance with the situation

APPENDIX D
POST-STUDY SYSTEM USABILITY QUESTIONNAIRE

**The Post-Study Usability Questionnaire
Version 3**

**Strongly
agree**

**Strongly
disagree**

			1	2	3	4	5	6	7	NA
1	Overall, I am satisfied with how easy it is to use this system.		☐	☐	☐	☐	☐	☐	☐	☐
2	It was simple to use this system.		☐	☐	☐	☐	☐	☐	☐	☐
3	I was able to complete the tasks and scenarios quickly using this system.		☐	☐	☐	☐	☐	☐	☐	☐
4	I felt comfortable using this system.		☐	☐	☐	☐	☐	☐	☐	☐
5	It was easy to learn to use this system.		☐	☐	☐	☐	☐	☐	☐	☐
6	I believe I could become productive quickly using this system.		☐	☐	☐	☐	☐	☐	☐	☐
7	The system gave error messages that clearly told me how to fix problems.		☐	☐	☐	☐	☐	☐	☐	☐
8	Whenever I made a mistake using the system, I could recover easily and quickly.		☐	☐	☐	☐	☐	☐	☐	☐
9	The information (such as online help, on-screen messages and other documentation) provided with this system was clear.		☐	☐	☐	☐	☐	☐	☐	☐
10	It was easy to find the information I needed.		☐	☐	☐	☐	☐	☐	☐	☐
11	The information was effective in helping me complete the tasks and scenarios.		☐	☐	☐	☐	☐	☐	☐	☐
12	The organization of information on the system screens was clear.		☐	☐	☐	☐	☐	☐	☐	☐
13	The interface* of this system was pleasant.		☐	☐	☐	☐	☐	☐	☐	☐
14	I liked using the interface of this system.		☐	☐	☐	☐	☐	☐	☐	☐
15	This system has all the functions and capabilities I expect it to have.		☐	☐	☐	☐	☐	☐	☐	☐
16	Overall, I am satisfied with this system.		☐	☐	☐	☐	☐	☐	☐	☐

**The "interface" includes those items that you use to interact with the system. For example, some components of the interface are the keyboard, the mouse, the microphone, and the screens (including their graphics and language).*

APPENDIX E
CONDENSED 14-ITEM SCHAEFER TRUST IN AUTOMATION SUBSCALE

What % of the time will <u>THIS</u> autonomous system...	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Act consistently	o	o	o	o	o	o	o	o	o	o	o
Function successfully	o	o	o	o	o	o	o	o	o	o	o
Malfunction	o	o	o	o	o	o	o	o	o	o	o
Have errors	o	o	o	o	o	o	o	o	o	o	o
Provide feedback	o	o	o	o	o	o	o	o	o	o	o
Meet the needs of the mission	o	o	o	o	o	o	o	o	o	o	o
Provide appropriate information	o	o	o	o	o	o	o	o	o	o	o
Communicate with people	o	o	o	o	o	o	o	o	o	o	o
Perform exactly as instructed	o	o	o	o	o	o	o	o	o	o	o
Follow directions	o	o	o	o	o	o	o	o	o	o	o
Be dependable	o	o	o	o	o	o	o	o	o	o	o
Be reliable	o	o	o	o	o	o	o	o	o	o	o
Be unresponsive	o	o	o	o	o	o	o	o	o	o	o
Be predictable	o	o	o	o	o	o	o	o	o	o	o

APPENDIX F
GENERAL TRUST IN AUTOMATION INVENTORY

Generally...	Disagree strongly	Disagree moderately	Disagree a little	Neither agree nor disagree	Agree a little	Agree moderately	Agree strongly
I can rely on automation to ensure my performance	o	o	o	o	o	o	o
Automation improves my performance	o	o	o	o	o	o	o
It is easy to follow what automation does	o	o	o	o	o	o	o
Automation makes use of all the knowledge and information available to produce its solution to the problem	o	o	o	o	o	o	o
I am confident in automation	o	o	o	o	o	o	o
Automation correctly uses the information I provided	o	o	o	o	o	o	o
The advice automation produces is as good as that which a highly competent person could produce	o	o	o	o	o	o	o
Automation always proved the advice I require to make my decision	o	o	o	o	o	o	o
Using automation increases my productivity	o	o	o	o	o	o	o
Using automation enables me to accomplish tasks more quickly	o	o	o	o	o	o	o

APPENDIX G
DEMOGRAPHICS SURVEY

Pilot # _____ Date _____ Grade / Rank: _____ Active Duty? Yes ___ No ___
 Service / Unit: _____ Age: _____ Gender: Male ___ Female ___
 Current Position _____
 Handedness: Left ___ Right ___ Glasses? Yes ___ No ___ Hearing Aids? Yes ___ No ___
 Highest Level of Education (Degree/Subject) _____
 Aircraft Qualifications (Hours) (e.g., AH-64D = 850 hours, OH-58D = 120 hours) _____

Flight Hours

Total Flight Hours	Fixed-Wing Hours	Rotary-Wing Hours	Combat Hours	Pilot in Command Hours	Instructor Pilot Hours	Maintenance Pilot Hours	Experimental Test Pilot Hours	AMC Hours	Night Vision (NVG) Hours	Weather (IMC/IFR) Hours

Specific Command roles (e.g., Air Mission Commander, Air Weapons Lead) _____

Aviation specific qualifications (e.g., Instructor, Maintenance, XP, Safety, Tactics, WTI) _____

Do you have NTC, JRTC, or JMRC experience? Yes ___ No ___

If yes, how many hours in COIN Operations? _____ How many hours in Decisive Action/Near-Peer Ops? _____

Do you have Aerial Gunnery experience as an Army Aviator? Yes ___ No ___

Do you have Hellfire engagement experience? Yes ___ No ___ If yes, check all that apply:

Captive Flight Trainer Missiles ___ Live missiles in training ___ Live missiles in combat ___

Do you have experience calling for Indirect Fires from Artillery, in training or otherwise? Yes ___ No ___

Civilian Aviation experience (e.g., CFII, Examiner, ATP, Aircraft ratings) _____

Have you teamed with UAVs (e.g., Grey Eagle, Shadow, Reaper) during:

-Training: Yes ___ No ___ If Yes, approximate number of hours _____ Which UAV? _____

-Combat: Yes ___ No ___ If Yes, approximate number of hours _____ Which UAV? _____

If you teamed with UAVs in training or combat, which Level of Interoperability (LOI) did you use:

LOI 2 ___ LOI 3 ___ LOI 4 ___ Not Applicable ___ (check all that apply)

Were you an AMC during missions where you teamed with a UAV?

Yes ___ No ___ If Yes, number of hours _____

Have you performed duties in a Ground Control Station as a payload or air vehicle operator?

Yes ___ No ___

If Yes, describe your duties in the Ground Control Station and how many hours: _____

Total number of months deployed to combat _____

What countries/areas of operation have you flown in (dates)? _____

Video Game Experience? Yes _____ No _____

Select your Favorite(s) (1= favorite, blank = no exp): FPS____, Action____, RPG____, MMO____, Strategy____,
Rhythm____

Rank frequency of use (1=most frequent, 4=least freq): Mouse+Keyboard____, Controller____, VR____,
Mobile____

Younger than 18 yrs average # hours per week _____ *Older* than 18 yrs average # hours per
week _____

Have you ever used VR headsets (e.g., Oculus Rift, HTC Vive)? Yes _____ No _____

If yes, briefly describe your experience with VR headsets (e.g., just a demonstration, play video games,
used in training) _____

APPENDIX H
POST-VIGNETTE QUESTIONNAIRE

Vignette Questions

How confident are you that you secured the ingress route?

- ☐ Not at all confident
- ☐ Somewhat confident
- ☐ Confident
- ☐ Very confident
- ☐ Extremely confident

How confident are you that you secured the LZ?

- ☐ Not at all confident
- ☐ Somewhat confident
- ☐ Confident
- ☐ Very confident
- ☐ Extremely confident

What were the most difficult/challenging aspects of this vignette?

What mistakes might a novice/inexperienced AMC make in this vignette?

System/Automation Questions

Which system capabilities do you feel contributed the most to your ability to perform this particular vignette successfully?

Were any important system capabilities missing or lacking that you would have used in this vignette?

Did you find any of the interface elements difficult or confusing to use? Did anything not behave in a way you expected? What important information or function is not available in the interface?

Did you find the vignette, and the way you were instructed to execute it, to be realistic?

To what extent did the MUM-T capability contribute to mission success?

- ☐ none
- ☐ very little
- ☐ somewhat
- ☐ helpful
- ☐ critical (could not complete mission without it)

APPENDIX I
SYNERGISTIC UNMANNED MANNED INTELLIGENT TEAMING
PILOT INTERVIEW FEEDBACK

Pilot Design Recommendations

Background: The SUMIT program recently completed an approximately three year manned/unmanned teaming (MUM-T) experiment featuring an advanced cockpit design, autonomous unmanned aerial systems (UAS) behaviors, an aided target recognition system (ATR), and other design features. The program ran 39 subject pilots through the experiment executing a total of 186 mission vignettes in both a COIN and high threat tactical environment (IADS Breach). Pilots trained on and flew simulated missions in the cockpit and systems designs developed by three separate vendors, IST, Toyon, and GE Aviation. The SUMIT research was hosted in the Future Open Rotorcraft Cockpit Environment (FORCE) flight simulator located at Intrepid Inc. in Huntsville, AL.

During the execution of the vignettes, subject pilots acted as the air mission commander (AMC) teamed with between 2 and 8 armed and unarmed UAS assets. Each pilot flew 2 practice and 4 data collection missions. Structured interviews were conducted following completion of one practice and all four data collection vignettes, with a final post experiment interview to capture more detailed data in specific areas of interest.

The purpose was to determine how to effectively measure system and cockpit design performance in a teamed tactical environment. A secondary purpose was to determine what cockpit design features and platform technologies would best expand the mission effectiveness of a single pilot, conducting a complex mission, controlling multiple ISR and armed UAS platforms. The post-mission pilot interviews were recorded and later transcribed into a Word document that captured their experiences with the SUMIT vendor designs during mission execution. All of the subject pilots had one of more combat tours flying attack, reconnaissance, or utility helicopters. The pilot comments voiced what they liked and did not like about the vendor design, often suggesting design improvements that would increase situational awareness and overall mission effectiveness.

Methodology. Review the recorded pilot comments and extract from them the stated and implied cockpit design and aircraft system improvements pilots felt would increase mission effectiveness when executing advanced teaming combat missions with multiple unmanned assets. Document those recommended improvements in a design guide format.

Design Guidance. The cockpit and aircraft system design recommendations extracted from SUMIT pilot interviews follows. The words “must” and “should” appear throughout this document. These words should not be interpreted at design requirements but rather recommended design solutions. The word SHALL is reserved for expressing a requirement and does not appear in this document.

General design guidance.

- The overall system design must facilitate the crewmember spending the majority of the mission at the command decision making level (AMC), and spend less time at the detailed operator input level.
- The system should minimize the number of button presses or individual system inputs required to execute simple tasks. Up to 550-650 individual button presses were routinely recorded during a 30 minute mission. This is excessive workload especially when projected over a typical 2 hour mission.
- The system design must allow the crewmember to be able to quickly and easily reverse/cancel an input that was made in error.
- Design the system so that the most common activities such as sensor slew, sensor zoom, or move an asset from one location to another can be accomplished using more than one system input modality. Examples are joystick or touch and drag to move a sensor, or thumb wheel or touchscreen pinch to zoom a sensor, or voice command or touch to switch primary sensor views from one asset to another.
- The system design should avoid using bezel buttons as the primary input modality. Bezel buttons are 40 year old technology.
- Use of long strings of text should be minimized. Font size should be easily readable from the design eye point in the cockpit. Graphical displays are easier and faster to interpret than text.
- The system design should include artificial intelligence (AI) that is capable of analyzing where the teamed asset is in the mission profile, what decisions are needed at this point in time, and should be capable of automatically executing or recommending a set of alternatives that the crewmember can choose from.
- The system design should be such that the crewmember has good situational awareness during all mission phases. The design should be intuitive and not require piecing together multiple separate pieces of information from different sources to form a mental picture of the tactical situation and status of mission execution.
- If the crewmember makes an error while using the system, there should be an indication of what caused the error and how to correct it. As an example, if the system is only capable of processing and 8 digit grid coordinate and the crewmember enters either 6 or 10 and gets an error, the correct entry should be made clear.
- The ownship must have an autopilot capable of terrain, obstacle, and threat avoidance at terrain flight altitudes. This includes a hover position hold and auto-land at remote landing zones.
- The cockpit design should include flight controls and a selectable primary flight display that can be accessed by the left or front seat pilot in the event manual aircraft control is

desired or required from that crewstation. The FORCE simulator only had right seat flight controls and primary flight displays.

- All functions that the cockpit design and system is capable of should be executable from either cockpit crewstation assuming a two pilot cockpit.

Digital map.

- The digital map must have grid lines for use as a position reference. The grid lines should have a user selectable declutter (on/off) option.
- The interface design should include a capability to digitally transmit a grid coordinate of a position on the map such as a sensor stare point without manual data entry (click and send without manually typing the coordinate).
- Keep labels of entities on the map as simple and short as possible to avoid excessive clutter. Consider labels where the detailed information relating to that object will only appear when a mouse pointer or finger touch on a touchscreen opens an information box.
- Keep the line size of map graphics to a minimum to avoid visual clutter.
- A flashlight type symbology or similar design solution should be used to indicate precisely where on the ground a teamed asset sensor is looking. This should include an outlined area that indicates the area on the map that the sensor view is covering. There should be a clear indication using color coding or other method of which sensor symbol on the map is being generated by which teamed asset.
- The design must facilitate a visually intuitive correlation between the sensor and the map to minimize the crewmember looking back and forth between the sensor and map displays to understand what is being looked at and where. Adding map graphics to the sensor view is an example of a design solution.
- The crewmember should be able to touch a teamed asset and drag it to a point on the map where the crewmember wants the asset to relocate. The asset should then fly to that new location on the map without further commands.
- The map should have a scale or some other method of measuring distances.
- Ownship heading information should be displayed on the map as well as the ownship sensor view.
- North-up or track-up map orientation should be selectable by the crewmember.
- The map must have an easy to use tool for creating map graphics. As an example, the Army Air Mission Planning System (AMPS) in use in the field today uses Falcon View which has robust battlefield graphics tools.

- If a teamed asset flies off the map the operator should be given an indication that it occurred and be provided a direction and distance where the asset is currently located.
- A north seeking arrow or other indicator of north should be displayed on the map.
- A telestrator (touchscreen whiteboard) capability should be included in the map design. It would be valuable in increasing input speed, increasing SA, and reducing button pushing workload. Drawing routes or areas with the finger on a touchscreen is technology that exists today. The crewmember should be able to command a teamed asset to follow the route or search the area inside the pattern drawn. There should be a capability to send the telestrator information digitally to another teammate, the TOC, the ground commander, or whoever might need the information.
- Teamed assets should be color coded or be clearly identifiable on the map at a glance. This design feature will enable the crewmember to quickly distinguish and identify both the aircraft position and movement as well as its associated sensor movement and stare point.
- The map should include a declutter capability. Icons can quickly group up and begin to overlap each other making it difficult to understand what is at that location. Selectively decluttering the map is important to SA. Once decluttered there must be an easy way to reverse the declutter process.
- The map should include color shading or other method to indicate where line of sight from a teamed asset or ownship and a threat weapon exists and does not exist. This design feature may allow a teamed asset to operate within the maximum effectiveness of a weapon system but not within line of sight of the weapon. This is a design feature of the Harris map in the E model Apache and has proven to be most valuable when operating at terrain flight altitudes.
- The system must facilitate labeling of previously unknown entities that were identified but not classified by the ATR system.
- The crewmember should be able to move the map using touch and drag or zoom in on the map using the 2 finger pinch, joy stick, thumb wheel, or other input method.

Sensor design.

- The optimum sensor should be gimbaled and capable of 360° slew in azimuth, and a minimum 30° up elevation and 180° down (looking backward) elevation. There should be no uncommanded sensor rotation at the nadir position.
- The sensor zoom function should be continuous not stepped in all sensor modes, with no perceptible latency between input and sensor reaction.
- Sensor zoom input should be either a pinching motion on a touchscreen display or by a thumb switch, short joystick, or other method. More than one method should be available.

- The sensor should be capable of streaming EO, FLIR, MTI, or SAR images (does not include ALE sensors). An option for a blended sensor image is desired. The crewmember should be able to switch between sensor modes with little effort using more than one input method (examples button, switch, touch, voice).
- Teamed UAS sensors (not including ALE) should be capable of identifying an enemy combatant holding a weapon at ranges well beyond the maximum effective range of onboard weapons such as Hellfire missiles. The longer range sensor capability will also allow higher mission altitudes for non-weaponized teamed assets such as Shadow, and will decrease noise signature detection and increase survivability.
- The ownship sensor should be capable of identifying an enemy combatant holding a weapon at 20 km or greater.
- The crewmember should be able to drive the ownship sensor or the sensor of any teamed asset to a previously stored target using a touchscreen input or other method (joystick, eye gaze, voice, or other).
- The crewmember should have the ability to command sensors from one or more teamed assets to stare at the same point or target with minimal effort.
- The crewmember must be able to manually track a moving target or hold the sensor cross hairs on a stationary target or object on the ground or in the air with minimal effort and a high degree of accuracy while the host aircraft is maneuvering.
- Live sensor feeds from the ownship and a minimum of 4 UAS assets (desired all sensors) should be simultaneously viewable by the crewmember.
- Sensor feeds should be color coded or other method to clearly identify which sensor feed is associated with which teamed asset. If color coding is used, there must be an easily recognizable difference in the colors. Subtle differences will not work as a design solution.
- Recommend one large (primary) and multiple smaller (minimized) live sensor feeds with a quick method of switching between the larger sensor view and the minimized sensor views such as a touchscreen tap, voice command, eye gaze, or other method.
- Earth referenced graphics and symbology should appear in the live sensor feeds. As an example, an NAI should appear as a graphically generated symbol on both the digital map and also in the sensor feed. During a route recon, a line representing the route to be searched should appear in the sensor feed as well as on the digital map. A ROZ should be depicted in the sensor feed as a colored 3D object or other method and on the digital map.
- If a teamed asset's sensor is performing an autonomous search, the crewmember should be able to temporarily take manual control of the sensor to look at specific objects or areas of interest then return the sensor to the automated behavior. The sensor should resume the automated behavior from the point where manual control was taken.

- All sensors in the team must have an image autotrack (IAT) capability.
- There should be no sensor jitter in any mode of operation or under any circumstance to include during zoomed-in maneuvering flight.
- There should be no perceptible latency between the crewmember input and the sensor movement during panning or zooming.

Commanding very small, precise sensor movements should be easy even when the sensor is zoomed in and the aircraft is maneuvering.

- Target labels applied in the map view should show up in the sensor view.
- The crewmember should be able to mark and label a target from the sensor view and not have to switch to the map view to attach a label. If a label is attached in the sensor view it should be repeated in the map view.
- There must be an indication on the digital map where the sensor has scanned during either manual or autonomous scanning. This could be a translucent color overlay or similar design to indicate, at a glance, what areas have been looked at and what areas have not been looked at. It should only “paint” areas where the zoom level is sufficient to detect human or vehicle targets. There must be a capability to set a decay time for the “painted” areas to indicate when the information is becoming stale.
- The design must facilitate a seamless correlation between the sensor and the map to minimize the workload and time delays associated with the back and forth between the sensor and map displays. Map labels and graphics that also appear in the sensor view is the type of design solution that will help mitigate this issue.
- There should be an indicator in the sensor view that shows the current percent of total zoom and the percent zoom remaining.
- If there are sensor gimbal limits, there should be a field or regard box or other indication showing where the sensor currently is with respect to those limits.
- Range from the sensor to the sensor cross hairs stare point should be displayed in the sensor view along with sensor azimuth or bearing.
- The amount of sensor movement with crewmember input should automatically adjust as the sensor zoom level is increased. As sensor zoom increases, the sensor movement should decrease. If not, the sensor will be too sensitive at higher zoom levels and will not be controllable.
- All teamed sensors need a single input stow capability. The stow position in azimuth should be along the aircraft centerline and level in elevation with respect to the aircraft fuselage.

- The crewmember needs the ability to slew the sensor to a grid coordinate that was received digitally from another source. Cut and paste is preferred over a manual grid coordinate entry.
- Active sensor views from all teamed assets should be displayed if there is space available. The size of the displayed sensor views should be adjustable but the minimum adjusted size should be large enough that the information presented is usable.
- The ability to switch primary sensor views should require only a single input such as a touchscreen input, voice command, eye gaze or other method.
- It is easier to zoom a sensor with a manual input such as a thumb switch or joystick than with pinch to zoom. However, both methods should be incorporated in the system design.

Autonomous behaviors.

- As an autonomous behavior is executing its assigned mission, the route that will be flown should be presented graphically to the crewmember and the time remaining to complete the autonomous behavior should be presented both in an analog presentation and in a digital time remaining presentation.
- As a minimum, autonomous behaviors should include route recon, area recon, named areas of interest (NAI) or points of interest (POI) recon, and target engagement.
- The crewmember must have the capability to take manual control of an asset that is executing an autonomous behavior and quickly move its location when it is receiving hostile fire or is flying inside the engagement range of a threat weapon.
- The crewmember must have the ability to temporarily take manual control of a sensor that is executing an automated search, use it, and then return the sensor to its previous assigned search behavior.
- When an autonomous behavior is selected, there must be a clear indication that it is in progress and the time remaining to complete.
- If an autonomous behavior fails to execute or is interrupted by loss link or other technical issue, the operator should be immediately alerted to the interruption or mission stoppage.
- If an autonomous behavior completes its assignment, the teamed asset should loiter in place until it receives a follow-on assignment. The operator must be alerted to the fact that the mission has ended and the asset is available for reassignment or available to repeat that mission. There should be a minimum of two different alerting modes such as an audio alert and a visual alert. If the desire is to repeat the mission, a single input is all that should be required. If reassigned to a different mission, the current values for altitude and airspeed should not have to be reset if they remain the same as the previous mission.

- The crewmember must be able to modify an autonomous behavior both before it executes and while it is executing without having to cancel the mission and reprogram a new, modified behavior.
- When executing an autonomous mission that ends in a loiter pattern, the operator should have the option to select a loiter position and stare point for the asset. Terrain, noise, enemy position all dictate the need be able to select an appropriate UAS holding position and flight pattern that are tactically sound.
- When executing an autonomous mission to recon an NAI, POI, route, or target on the ground, the UAS asset should not fly directly over the intended stare point on the ground. The automatically or manually selected loiter position or route recon position should be offset from the stare point by a user selected distance or a minimum default distance of 1.5 km.
- If the operator programs an autonomous NAI or POI recon, there should be an option to select the type flight pattern to be flown at the loiter position such as race track, figure 8, circle, etc., and the lengths of the pattern legs should be programmable with an option to select a default size if a specific size is not required.
- Once an autonomous behavior has completed the assigned mission, there should be an automatically generated summary of what was found during the mission. This summary could easily and quickly be provided as a SITREP to those who need the information.
- The system design must facilitate being able to link consecutive autonomous tasks. When a task is complete, the next task should begin automatically or at a scheduled time without additional crewmember input. There must be a task status notification that a task has ended and a new task has begun.
- During the enroute portion of an autonomous behavior, the air asset should automatically route around existing restricted operating zones (ROZ) and known threat locations.
- If color is used to distinguish assets on the digital map, the color assigned to that asset should remain unchanged when the asset is assigned to a new task or mission.
- The system must be able to assign multiple teamed assets to a single mission such as an area or route recon. The assigned assets should automatically divide up the task into equal portions and execute it without the crewmember having to program each asset individually.
- If multiple teamed assets are carrying out a single mission such as route recon, and one asset or more assets are removed from the autonomous behavior for any reason, the remaining assets should automatically reassess and pick up the departed assets part of the mission.
- When setting up an autonomous mission, the crewmember should have to select the teamed assets desired for the autonomous behavior not deselect the assets that were not desired to be included in the mission. The system should recommend which assets should

be involved based on mission efficiency factors (fuel status, proximity, reaction time, available weapons and sensors) and the crewmember can select from that list or select an asset that is not listed.

- The crewmember must have the ability to modify an autonomously generated route.
- Routes generated by the autonomous mission planner should be tactically sound. As an example, a flight route should be offset from the route that is under reconnaissance. The asset should enter the loiter pattern prior to flying directly over a target when the route ends in a loiter flight pattern.
- When there is an altitude change during execution of an autonomous maneuver the change in altitude should occur while enroute and not by spiraling up or down from the present position of the asset prior to proceeding on the route. Spiraling up or down in present position is not a realistic behavior and increases the time to get the asset on station and increases threat exposure.
- The crewmember should be able to specify which end of the route to begin an autonomous route recon behavior.
- The system should incorporate an automatic sense and avoid capability for airspace deconfliction with other teamed autonomous assets.
- If an autonomous behavior is modified or a new behavior selected, the current altitude and airspeed should be the default settings and not have to be re-entered unless the crewmember desires to change these values.
- Mission status information needs to be displayed at some logical location. Mission status information should include weapons and fuel status information, time and distance remaining to complete the autonomous behavior, primary flight information like altitude, airspeed, and heading information and other information that is critical to mission execution and situational understanding by the crewmember.
- There should be no default time limit for an automated task to execute.
- The system should be designed to require as few user inputs as possible to assign a mission task or autonomous behavior.
- The crewmember must be able to create an area on the map, including irregular shapes, using map graphics and command an asset to search inside the drawn area.
- The system design must be capable of allowing the crewmember to select roads only during an autonomous area recon.
- The system must be able to allow the crewmember to create priority search zones within the boundaries of an area recon.

- There must be a resume function that will allow the crewmember to resume the autonomous behavior after temporarily being interrupted by manual control.

Aided target recognition (ATR).

- An ATR system should be considered a mandatory capability when teamed with and controlling multiple UAS assets from the ownship.
- The ATR system must provide a dual alerting system such as a flashing of the sensor image and an audio alert or other alert when a potential threat has been detected.
- The ATR system should automatically engage image autotrack (IAT) to capture a detected moving or stationary target and track it until the operator provides input on how to process it such as store, track, track and store, or ignore. Targets that were selected to be ignored should not be alerted on by the ATR system again if the sensor passes over that same target.
- The ATR should be sensitive enough to discriminate between friendly, neutral, and enemy alerts and not alert on non-threat entities.
- The ATR should be capable of indicating to the crewmember the type target detected. Examples of the level of detail are individual combatant, armed light skinned vehicle (technical), armored vehicle (tank, BMP), or mobile or stationary anti-aircraft weapon system.
- The ATR must be capable of prioritizing a list of stored targets based on the threat to the ownship and teamed assets, threat to the ground force, or threat to overall mission accomplishment. As an example, the priority of threats to the ownship would first be air defense artillery (ADA), then armor, automatic weapons, followed by lesser threats.
- The ATR should be capable of automatically populating the digital map with the location of a detected threat, and labeling it appropriately with a MIL-STD 2525 icon or other label that is easily read and understood. The crewmember must be able to manually change the target ID label if the system generated label is incorrect or questionable.
- The system design should be capable of grouping and labeling targets that are clustered together such as a squad of enemy infantry that occupy a compact position on the ground or occupy a bunker, building, or other structure.
- The crewmember should be able to select the type of target being searched for. The minimum selection should be vehicles, individual combatants, or both. More detailed choices such as armored vehicles or armed individual combatants is a desired design feature.
- There must be a clear indication to the crewmember when the selected sensor zoom level is sufficient for the ATR to detect a target.

- When searching with the sensor using an autonomous mode, an automatic minimum zoom level should be set by the system that will coincide with the type of target the crewmember has chosen for detection by the ATR system such as an individual combatant or combat vehicle.
- When the ATR system sends an alert and there are multiple targets in the sensor view, it should be clear to the crewmember which target generated the alert. IAT lock-on is an example of a design solution.
- The ATR design must minimize the number of “unknown” target alerts. Slewing and zooming on each target in a list of unidentified targets to get a positive ID is time consuming and workload intensive.
- If the crewmember discovers a threat that was not picked up by the ATR system, there should be a capability to mark the target on the digital map and label it. An automatic update to Blue Force Tracker (BFT), or Link 16 battlefield SA should be sent.

Weapons engagement.

- The system design must show weapons engagement symbology that depicts as a minimum, the gun target line, danger no fly areas left and right of the weapon fly-out trajectory, optimal lasing positions for cooperative engagements, and symbology showing the minimum and maximum ranges for the selected weapon.
- The number of individual steps required to engage a target should be minimized to the maximum extent possible while preserving accuracy and safety. The crewmember must be able to get a shot off quickly to avoid losing a target.
- The system design must be able to quickly transition from an autonomous behavior such as an area search, to a weapons engagement once a target is found. The crewmember should not have to stop the autonomous behavior or cancel a current mission and program an engagement. The system must be capable of quickly and seamlessly transitioning to an engagement from any autonomous behavior with a minimum of crewmember input.
- The crewmember must be able to manually select the shooting and lasing air assets even when the capability to auto-select those assets exists in the system design.
- A weapons engagement should be automated to the point where shooting a target with a weapon does not require 100% of the operators attention for the entire launch sequence.
- Targets that have been identified, labeled, and stored for destruction should be listed in a prioritized format such as ADA and MANPADS first, armored vehicles second, and down the list.
- Missile flight time remaining to impact should be displayed following missile launch.

- The crewmember should be able to engage multiple targets in rapid succession without having to program subsequent engagements as a new mission. Targets such as multiple vehicles in a convoy sometimes require multiple missiles in flight with timed spacing between missile impacts. The crewmember simply moves the crosshairs from one vehicle to the next. This capability exists in the current aircraft fleet.
- The crewmember needs a cockpit indication when the location of friendly forces or civilian personnel are dangerously close (danger close) to the target being engaged to help prevent fratricide.
- Automated behaviors should be transparent, easy to understand, and predictable during execution. Trust in the automation is key to crewmembers using it. If it requires too many steps or button presses to set up or executes in a way that is not what the crewmember anticipated or executes in a manner that is not tactically sound, the crewmember will find work-arounds or not use the system entirely.
- When a target is destroyed, an “X” or some other symbol should appear over the target on the map that indicates target destroyed. Also, an automatic update to BFT or Link 16 should be sent to update the threat picture. If there was a threat ring around the target the ring should disappear when the target is destroyed.
- If the crewmember has difficulty executing an engagement the system should present information on what to do to correct the problem and get it to engage. As an example, if the shooting platform is out of range for the selected weapon, the system should notify the crewmember that excessive range is preventing the shot, or if there are no missiles remaining there should be a notification.
- Weapon system status should be presented to the crewmember. This includes, as a minimum, number and type of weapons remaining, type weapons fuse/warhead selected (HE, delayed fuse, sabot, fleshettes, others), firing mode (standby or armed), actively firing (laser).

Touchscreen display.

- Touchscreen displays must be located at a comfortable distance from the crewmember and not require an excessively long reach. The design should consider an angularly adjustable display that will allow a direct crewmember input to the left and right displays.
- A touchscreen capability should be considered for all cockpit displays that the crewmember interacts with. Touchscreen displays will work in helicopter cockpit environments and are in use today.
- Multi-touch is a required design feature to facilitate pinch to zoom, touch and move sensors and maps, and many other applications. The amount of reaction to a pinch input should be consistent or the same for all sensors being controlled such as a Shadow sensor and Gray Eagle sensor zoom reaction should be identical.

- The display must be touch activated. Displays that use an IR plane across the surface or other near touch method are not usable in a cockpit due to the excessive inadvertent activations.
- Touch points on the display must be large enough to accommodate some level of inaccuracy inherent in a vibrating helicopter environment. Also fatigue and excessive workload will degrade fine motor skill so this must be considered when sizing the touch points.
- A telestrator (touchscreen whiteboard) capability should be incorporated in the system design. A telestrator would be valuable in increasing input speed, increasing SA, and reducing button pushing workload. Drawing routes or areas with the finger on a touchscreen is technology that exists today. The crewmember should be able to command a teamed asset to follow the route or pattern drawn or execute a recon within the boundaries of a drawn shape. There should be a method of sending this telestrator information digitally to another teammate, the TOC, the ground commander or whoever might need the information.
- There should be no perceptible latency between the touch input and the system reaction.

Communications.

- The system should have one or more forms of digital communications that is easy and fast to use. Pre-formatted messages, speech to text, automatic addressing, auto-fill formats for standard reports such as a SITREP, group messaging, and many other methods could be used to make digital communications the preferred method over radio communications. Digital messaging must be secure.
- The system design must include an alerting system when new digital messages are received. The alert for incoming urgent digital messages should be different than routine messages. Urgent message alerts should be in a prominent place such as near the master caution light.
- A design feature that would allow the crewmember to click and drag information to a report format such as a SPOT report would save time and decrease workload.

Chat.

- The chat display must be close to the primary field of view. The FORCE cockpit had the chat display on the center console so when a message came in the crewmember had to look down to read it. During high workload chat messages were ignored.
- There should be a capability to cut and paste information into a chat message to avoid manual entry.
- When a critical message is received, there should be a special audio or other alerting method to indicate this is not routine chat.

- The system should provide for pre-formatted chat messages to reduce the time and workload required to generate reports like a SITREP or to respond to information requests that can be answered with a short reply like WICO or Roger.
- There should be a speech to text capability for creating chat messages.
- Chat messages should be color coded to indicate which messages are intended for which node on the chat network. As an example, if the ownship assigned color is green, then the crewmember can quickly see which chat messages were intended for the ownship.

Aircraft survivability equipment (ASE) (UAS aircraft do not typically have ASE)

- A hostile fire indicator should be included in the suite of ASE. The indicator should pinpoint the origin of the fire, the direction and range from the aircraft being engaged, type weapon (small arms, machine gun, crew served weapon), and should automatically populate the digital map with the threat location to include a label indicating the type of threat.
- The ownship ASE suite should include an IR warning detector, radar warning detector, laser warning detector, and hostile fire indicator as a minimum. An appropriate ASE suite should be installed on UAS assets if required for the mission and weight, space, and power limits can accommodate the required sensors.
- When the ASE system detects the air asset is being engaged by enemy fire, the sensor on the air vehicle being engaged should automatically slew to and lock on to the source of the fire except when that sensor is directly involved in a weapons engagement and a missile is in flight.
- A chat message should not be the sole means of alerting the crewmember that the ownship or other air asset is taking hostile fire.
- There must be an indicator that shows quantity of chaff, flares, or other expendables remaining.
- There should be an indication on the map like a threat ring, color shading, or other method showing the maximum effective range from a known threat weapons position.

Speech activated systems.

- Speech activated systems should be considered in the cockpit design. Speech to text and speech to activate will reduce pilot workload and provide a system input method other than manual input.
- Speech input should never be the only input method available to the crewmember.

Eye gaze activated systems.

- Eye gaze activated systems should be considered in the cockpit design. Eye activated systems have the potential to reduce pilot workload when properly designed and integrated. It could provide an alternative to manual input.
- Eye gaze input should never be the only input method available to the crewmember

Gesture activated systems.

- Gesture activated systems should be considered in the cockpit design. Pinching on a touchscreen display is a form of gesture control. Gesture control could provide an alternative to manual input.
- Gesture input should never be the only input method available to the crewmember

Conclusion. This design guide was written after reviewing the comments made by 31 combat experienced pilots (8 pilot interview audio recordings were unavailable) who flew 186 separate simulated MUM-T missions while controlling up to eight teamed UAS assets. The implied cockpit design and aircraft system improvements the pilots felt would increase mission effectiveness when executing advanced teaming combat missions were extracted from the pilot structured interview responses following each of the 186 vignettes. The responses were paraphrased in a design guide format, and sectioned into major functional areas.