# Mixed Initiative Team Performance Assessment System (MITPAS) For Training and Operation

**A. Freedy, E. De Visser, G. Weltman**
**Perceptronics Solutions, Inc.**
**Sherman Oaks, California**
**afreedy@percsolutions.com**
**edevisser@percsolutions.com**
**gweltman@percsolutions.com**

**Nicole Coeyman**
**U.S. Army RDECOM-STTC**
**Nicole.Coeyman@us.army.mil**

## ABSTRACT

Unmanned vehicles (UVs) are being developed and fielded at an unprecedented rate in various environments for both civilian and military purposes. Despite the term "unmanned," control of such vehicles requires considerable inputs from those operating the UVs and others in the C2 system. To optimize such mixed initiative teams we must develop new methodologies and measurements for evaluating and understanding human-robot team performance. The Mixed Initiative Team Performance Assessment System (MITPAS) provides such new methodology. MITPAS consists of models, tools and procedures, including an OneSAF-based simulation environment, with which to measure the performance of mixed manned and unmanned teams in both training and real world operational environments. This paper describes MITPAS and the results of several initial experiments conducted to validate the measures and gain insight into the effect of robot competence on overall human-robot team performance, operator trust and operator situational awareness. Our initial results are indicative of the type of new insights into human-robot team behavior that can be gained by combining the measurement power of MITPAS with realistic simulations of tactical UV operations. Consequently, we are working to make the readily-customized MITPAS available to other researchers and developers for use with their simulations, scenarios and special measures.

## ABOUT THE AUTHORS

**Amos Freedy** is President of Perceptronics Solutions Inc. He is a recognized expert in the field of AI-based robotics and in the planning and management of R&D programs involving a combination of engineering, bioengineering and behavioral disciplines. He is currently directing several U.S. Army and DARPA programs dealing with interactions between teams of intelligent robots and human and computer control systems. Dr. Freedy received a B.S. in Electrical Engineering, a M.S. in Bio-Engineering and a Ph.D. in Electronics and Man-Machine Systems from UCLA.

**Ewart De Visser** is a Human Factors Engineer at Perceptronics Solutions, Inc. His research interests include human-robot interaction, decision making, and neuroergonomics. He received a B.A. in Film Studies from the University of North Carolina at Wilmington and a M.A. in Human Factors and Applied Cognition from George Mason University. He is currently pursuing a Ph.D. degree in Human Factors and Applied Cognition at GMU.

**Gershon Weltman** is Principal Scientist at Perceptronics Solutions, Inc. His research interests include training, human factors engineering and user interface design. Dr. Weltman's professional experience is centered on his long-term executive and technical management of Perceptronics, Inc. At Perceptronics, Inc. Dr. Weltman directed the design, production and delivery of many innovative simulation and decision support systems including SIMNET, the system that introduced networked virtual simulation to the armed forces. Dr. Weltman received a B.S. in Engineering, and a M.S. and Ph.D. in Man-Machine Systems Engineering from UCLA.

**Nicole Coeyman** is a Science and Technology Manager for the Asymmetric Warfare Virtual Training Technologies program at the U.S. Army Research Development and Engineering Command (RDECOM) Simulation and Training Technology Center (STTC). Her research interests include robotics, simulation technologies, massively multi-player online game (MMOGs) technologies, and digital systems. She received a BS in computer engineering from the University of Central Florida, and is currently pursuing a Master's Degree in Industrial Engineering (Engineering Management) from the University of Central Florida.

# Mixed Initiative Team Performance Assessment System (MITPAS) For Training and Operation

A. **Freedy, E. De Visser, G. Weltman**
**Perceptronics Solutions, Inc.**
**Sherman Oaks, California**
**afreedy@percsolutions.com**
**edevisser@percsolutions.com**
**gweltman@percsolutions.com**

**Nicole Coeyman**
**U.S. Army RDECOM-STTC**
**Nicole.Coeyman@us.army.mil**

## INTRODUCTION

Unmanned vehicles (UVs) are being developed and fielded at an unprecedented rate in various aerial, ground, and underwater environments, for both civilian and military purposes. Despite the term "unmanned," control of such robotic vehicles requires considerable manpower from those operating the UVs, users of the information provided by UVs and command and control personnel. Mandates to reduce manning in the military have led to initiatives to distribute control of multiple heterogeneous UVs, to a small number of human operators. It is widely recognized, therefore, that in the future teams of human operators will need to supervise and control a much larger number of UVs. The goal of such "mixed initiative" human-robot teams is to extend manned capabilities and act as "force multipliers", as in the US Army Future Combat System (Cosenzo et al, 2006; Barnes, Parasuraman & Cosenzo, in press).

The inevitability of mixed initiative introduces a new and unique aspect to the psychology of team performance: that is the interaction of two cognitive systems -- human and autonomous unmanned robot. In addition to the critical performance factors traditionally associated with human teams -- which include information exchange, communication, supporting behavior and team leadership -- the mixed manned/unmanned team adds a number of challenging new dimensions. Foremost among these is the ability of the human team to manage, predict, collaborate and develop trust with unmanned systems that may sometimes exhibit fuzzy responses in unstructured and unpredictable environments.

If the vision of coordinated teams consisting of humans and robots is to be realized, it is essential to understand how small teams of operators can effectively control a large number of UVs of different types and with different capabilities. This requires analyzing the issues related to levels of automation (Parasuraman, Sheridan, & Wickens, 2000), human-robot interfaces (Adams & Skubic, 2005), and robot autonomy. It also requires developing new and effective methodologies for evaluating human-robot team performance. Such new methodologies are critical both for training and operational purposes.

Our R&D challenge has been to develop a new simulation environment in which to study mixed initiative behavior *and* a set of system-specific measures of behavior on which to base assessment of the mixed initiative team performance. As reported by Freedy et al (2004) we developed a performance model for mixed initiative tasks in the first phase of our work. A brief summary of the performance model will be given here. Subsequently, we have conducted several experimental studies that have provided initial insights into the performance of mixed initiative teams as well as into the viability and utility of our measures. Results of these studies will also be summarized and discussed here, following a brief description of our simulation environment and experimental methodology.

The performance model represents a particular challenge since the measures must be unique to the information and decision environment associated with human-robotic teams, and must directly link together behavioral processes important to mixed manned/ unmanned tactical outcomes. The measures need to provide feedback for skill improvement in collaboration as well as adaptation to stress and workload, and they should ideally help define the training needs themselves.

## PERFORMANCE MODEL

In accord with these goals, we have created a System Performance Model that captures the critical performance attributes and distinct processes of the behavior composition environment. Our objective has been to identify the dimensions of performance that contribute to effective outcomes of collaborative manned-unmanned tasks, and in particular to formulate measures that are unique to the collective teams of humans and robots. Accordingly, we have built a taxonomy of specific processes which can be decomposed into explicit behavioral objectives side-by-side with measures of effectiveness based on actual outcomes. Our focus is on process measures that are closely linked to outcomes, because it is these measures that will provide the feedback necessary for training. The three levels of team processes critical to training evaluation and remediation are: (1) individual human; (2) team human; and (3) collective human/robot team.

We decomposed the processes into these three levels and developed taxonomy of measures for each level. We narrowed the performance measures to the simplest factor structure that adequately cover the dimension of teamwork as was found by previous investigators (Barnes et al, 2002) The actual Performance Model consists of a multi-dimensional task process performance schema which (1) aggregates the performance measures at each level, (2) provides for training feedback at each level, and (3) provides a multi-attribute discriminate function to determine an overall level of proficiency as well as a "pass-fail' score. The weights of the attributes will be established in simulations in which the linkage between specific task performance measures and outcomes can be estimated.

There are two main types of measures: Measures of Performance (MOP) and Measures of Effectiveness (MOE); these are defined separately below.

***Measures of Performance (MOP).*** These are observable and derived measures of the operators' task skills, strategies, steps or procedures used to accomplish the task. They consist of the cognitive and interactive processes of the individual and team in collaborating together and controlling the robotic entities in a coordinate manner. MOP evaluates the human factor involved in a complex system. MOP was divided into 3 distinct classes of processes dimensions:

- **Human Team Processes** - These processes represent the dimensions of the human team interaction
- **UV Management and Control Processes** - These processes represent the tasks associated with real time control and monitoring of the autonomous entities
- **Human/Robot Team Processes** - These processes represent the dimensions of the human interaction with the robotic elements

***Measures of Effectiveness (MOE).*** These measure the "goodness' of the composed behavior in quality and the execution of war-fighting tasks. MOEs are influenced by much more than human performance. These measures also contain variance accounted for by system design, the surrounding environment and luck. The measure consists of the following dimensions

- **Mission Effectiveness** - Observable measures of the success of the mission as determined by objective military criteria.
- **Behavioral Effectiveness** - Measures of the dimension of behavioral effectiveness of the system in the battlefield

In addition, we have adapted for our mixed initiative studies other conventional psychological measures of performance, for example, the measures of situational awareness described below.

## MITPAS ENVIRONMENT

MITPAS is a complete simulation environment for performing experiments designed to measure and assess the performance of mixed human-robot teams in a variety of military and non-military situations. As shown in Figure 1, the total MITPAS environment combines several novel components:

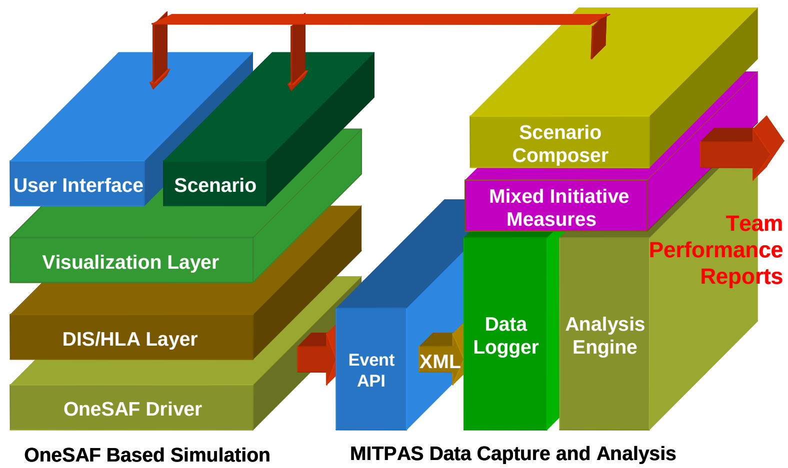


**Figure 1 MITPAS Components**

1. An OneSAF Based Simulation of a tactical environment and robotic command and control system. OneSAF is a generalized set of models and tools designed to supply semi-autonomous entities to a variety of simulations. Our original MITPAS used OneSAF OTB, our current system uses OneSAF OS (OOS).
2. A Data Capture and Analysis system consisting of tools and procedures to measure the performance of mixed manned and unmanned teams in both training and real world operational environments. It includes API, data logging and analysis components that can be applied to a wide variety of simulation environments.
3. A set of event-based test scenarios characteristic of anticipated military operations. Our event-based scenarios reflect mid-term plans for DOD's Future Combat System (FCS) as well as current uses of robotic systems in combat.

## COMMAND AND CONTROL SIMULATION

Figure 2 shows the command and control configuration for our simulation environment and experimental studies.

The Battlemaster, who may also play the platoon leader (see below), is in charge of the experimental procedures, the progress of the scenario, and communication with the Unmanned Ground Vehicle controller who is the actual experimental participant. The Unmanned Air Vehicle (UAV) controller has been a virtual participant to date.

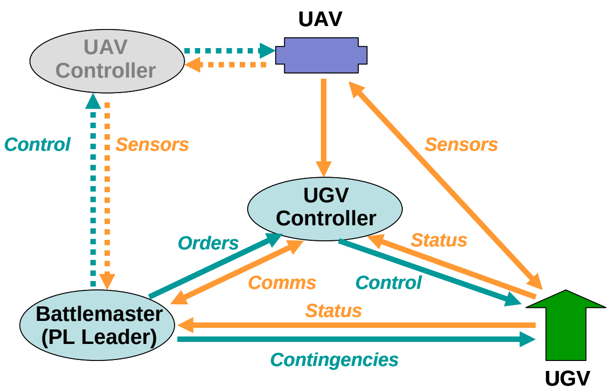


**Figure 2. Command and Control Configuration**

The command and control stations used by the Battlemaster and the UGV controller are very similar. This station as shown in Figure 3 is comprised of a tactical situation map, a 3D simulation of the environment, and a UGV status and communication display represented on three 19 inch monitors.

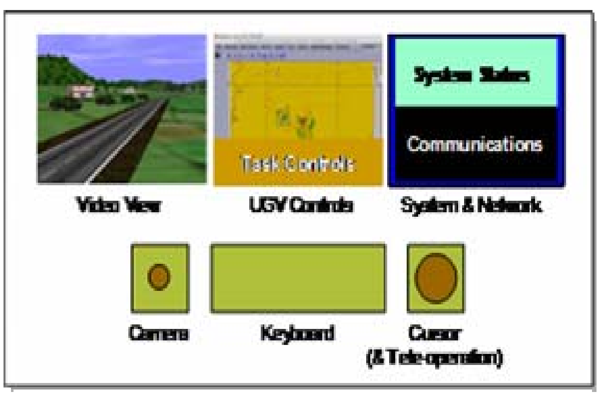


**Figure 3. UGV Controller Station**

The operator uses a keyboard, track-ball mouse, and joystick to control the system. The tactical situation map, generated by the military simulator OneSAF 2.5 on the center monitor, shows the UGV's geographic position and also gives an overview of the entire tactical situation.
The UGV control interface and the map control features enable the operator to assign the UGV specific tasks within the mapped area. A 3D representation of this environment, simulated by MAK Stealth 5.4 display on the left screen, can show live action from the viewpoint of a camera mounted on the UGV. The operator can tele-operate the UGV using a joystick in combination with gas and brake pedals if this becomes necessary.

The right monitor shows a java-coded UGV status and communications display. The status display provides real-time information to the operator about the UGV, including its direction, current target, surrounding friendly and enemy units, supply levels, and any vehicle malfunctions. Communications between the operator and the Battlemaster are facilitated through instant-message capabilities.

## TACTICAL SCENARIO

We have studied collaborative human-robot behavior in a scenario that took place at a simulated location with typical tactical features of roads, forests and a small village or built-up area, as shown in Figure 4.

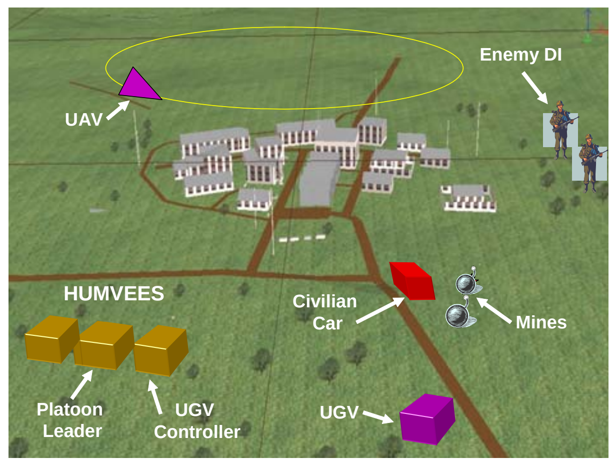


**Figure 4. MITPAS Experimental Environment**

The operator controlled an unmanned ground vehicle (UGV) as part of a reconnaissance platoon, whose mission was to ensure that a route through the area was safe for passage by eliminating all surrounding enemies. To do this, the operator had to move the UGV to a checkpoint where it could commence targeting an enemy, firing on the targeted enemy, and repeating this process for each of 6 attacking enemies. Operators also had to monitor and evaluate the autonomous targeting and firing capabilities of the UGV and take over control of the vehicle if these autonomous processes would cause a delay or fail completely.

The UGV could be set to one of three levels of targeting and firing competency – low, medium and high. Operators were trained to recognize the distinct symptoms associated with competency failures in autonomous targeting and UGV firing. Operators could improve targeting by either manually selecting an enemy on the battlefield or rotating the UGV manually to face the enemy. Firing competency could also be enhanced by moving the UGV closer to the enemy.

Participants were instructed to perform an override only when they thought the overall mission time would decrease as a result of their intervention. A secondary task included reporting each enemy kill, target overrides, and check point arrivals to the Battlemaster. Operators completed the mission by arriving at a second check point.

## METHODOLOGY

Twelve young adults (4 females and 8 males) varying in age from 18 to 25 served as UGV Controllers. Most participants had several years of gaming experience. Participants first received about 1.5 hours of training on proper use of the UGV Controller Station by completing exercises using a training manual.

Participants then completed 5 trials of 3 levels of UGV Firing Competency behavior (low, medium and high) totaling 15 trials. Fifteen unique scenarios were created varying the locations of the enemy to ensure that participants built up unique situation awareness during each mission. At about midway in the trial subjects responded to a Situation Awareness Global Assessment Technique (SAGAT) question (Endsley, 2000), and after each trial participants filled out the NASA Task Load Index (TLX) for subjective workload and an adapted version of a common subjective trust and self-confidence measure used in earlier automation research (Lee & Moray, 1992).

In the following we describe the experimental results in three categories – general performance, trust, and situational awareness during the course of the trial.

## GENERAL PERFORMANCE

We examined the effects of different levels of UGV Firing Competency and Competency Order on several dependent variables such as overall mission time, manual control distance, and workload. All collected data were submitted to a 3x5x6 repeated measures analysis of variance (ANOVA) with UGV Firing Competency (high, medium, low) and Trial (1 through 5) as the two within-subject factors and the six possible orders of UGV Firing Competency as a between-subjects factor.

Figure 5 below shows the results for (a) Mission Time vs. Trial (top left), (b) Distance Traveled in Manual Control vs. Competency Order (bottom left), (c) Trust vs. Trial (top right), and (d) Workload vs. Trial (bottom right).

Overall mission time increased as UV competency decreased, $F(1,8) = 3.83$, $p < 0.05$. In the first trial of the low competency condition, the average mission time is relatively high compared to the high and medium levels of competency. This is not surprising as the robot would almost always miss enemies in the low competency condition unless the participant intervened. Interestingly, as the trials progress overall mission time decreases for the low competency condition. One logical explanation for this effect is that as participants experience the low competency level of the UV, they learn to compensate by moving the machine closer to the enemy targets and therefore increase its firing accuracy, resulting in a shorter overall mission time. This effect is also confirmed by the decreasing average time it took to kill an enemy across trials in the low competency condition, $F(1,8) = 3.68$, $p < 0.05$.

The subjective workload measure supports the observed decrease in mission times for low competency levels across trials. High workload in the low competency condition decreases as trials progress, finally reaching the levels of workload in the medium and high competency conditions, $F(1,8) = 4.90$, $p < 0.001$. This effect is consistent with the idea that participants become more capable of handling the unreliable UV in manual control.

The effects of Firing Competency and Trial on workload and overall mission times provide converging evidence that the operator learns to adjust to the UVs performance in order to reduce overall mission times by overriding the UV's autonomous capabilities and engaging in manual control. An interesting sidelight of the operators' adjustment was that while there was improvement in performance for the low competency condition, there was virtually none for the medium competency case. It appeared that operators adjusted better to a UV with distinct behavior characteristics than to one with more indeterminate behavior.

The effects of Competency Order on the distance traveled in manual control show that manual control by operators differs based on which level of competency was presented first. When high competency was presented first, manual control tended to be low in the medium and high competency conditions, but when low competency was presented first, manual control was higher in the high and medium competency conditions, $F(1,10) = 8.68$, $p < 0.05$. This indicates that initial experience of low robot competency makes the operators insensitive to subsequent improvements in competency.

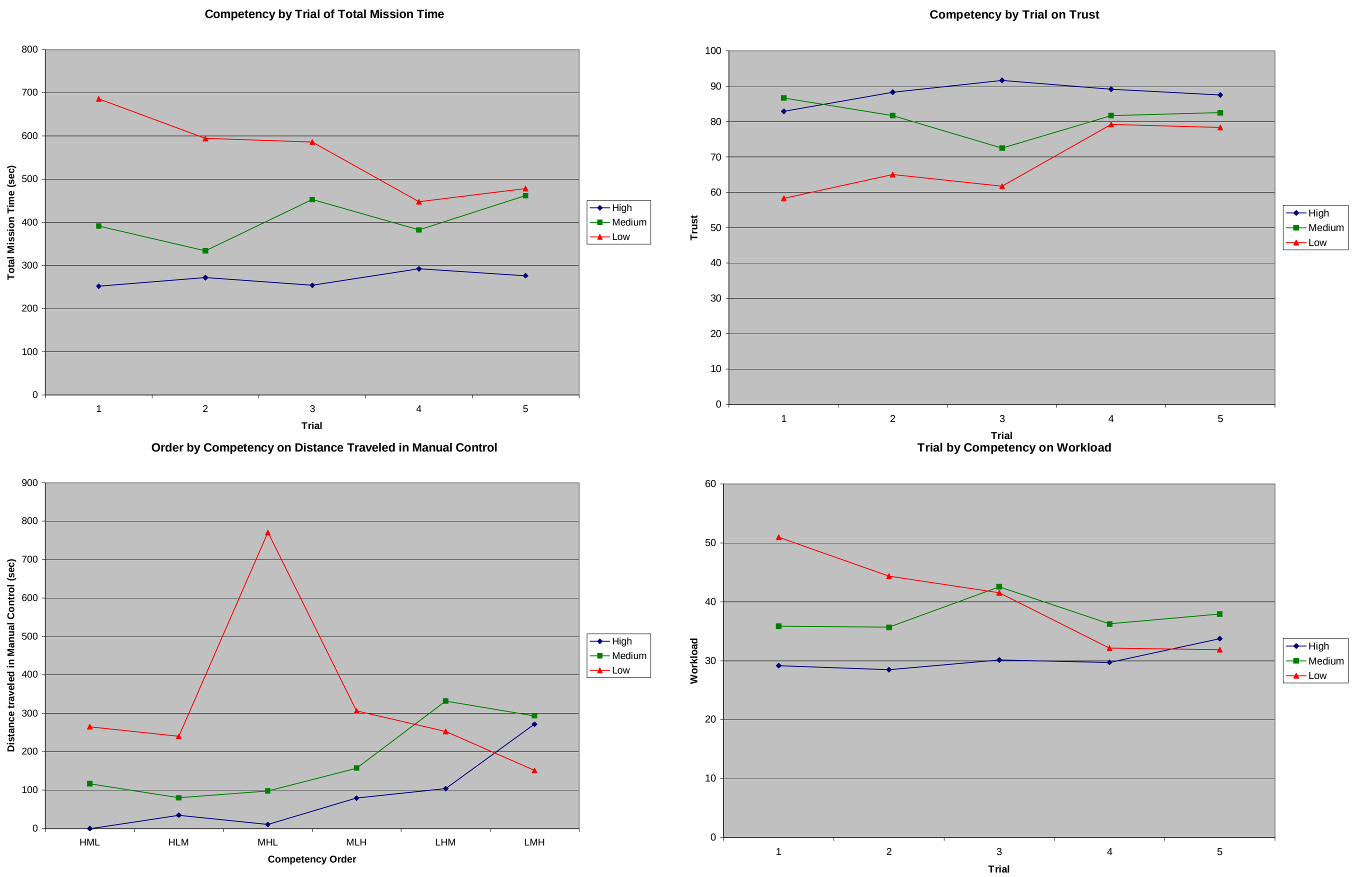


**Figure 5. General Performance Results**

## TRUST RESPONSE

### Background

Trust in automation is not a new concept and has been researched considerably (e.g. Lee & Moray, 1992; Parasuraman & Riley, 1997). Similarly to the significance of trust in human-human teams, trust in human-robot or *mixed initiative* teams is a key factor in determining the success of such a team. However, a mixed initiative team also creates unique challenges for developing trust because the robotic element may not be able to convey the necessary cues to develop trust as humans would among one another.

Lee and See (2004) propose that humans develop trust in a combination of three interplaying processes: analytic, analog, and affective methods. The analytic process is a rational approach to trust and assumes that when humans make decisions in uncertainty they use a cost-benefit analysis to determine appropriate trust for the system. If, for example, a robot makes a potentially costly error the operator may be less trusting and take over control to avoid future automation mistakes.

Methods for analyzing trust involve categorizing observed characteristics of the system and generalizing them to a broader set of assumptions. Applied to the mixed initiative team, such assumptions could be formed based on experiences with a certain type of robot that can do a particular task well. Finally, trust also forms purely affectively. Emotions may even supersede rational thinking and therefore can play an important role in relying on automation.

### Subjective Measurement of Trust

Figure 5c (upper right) shows the effect of robot competency on subjectively measured trust over the set of 5 trials for each competency condition.

The results show that overall, subjective trust correlates with the overall competency of the UGV -- as might be expected. But the overall result should make us reassess what “trust” actually means to UV operators. That is, our results indicate that subjective trust increases about 35% from 58 to 78 over 5 trials for the low competence UGV, while for the medium and high competence robots trust remains relatively unchanged at levels of 85 and 90, respectively, $F(1,8) = 4.68$, $p < 0.001$.

It appears that trust concerns not just an expectation of *correct performance*, but also an expectation of *level of performance*. Once the operator knows that the UGV will make mistakes, he or she *trusts* it to make those mistakes. Because the operator of the medium competency level UGV can not seem to get a good handle on what it will do, the subjective Trust level actually *decreases* slightly over 5 trials.

### Objective Measurement of Trust

Creation of an objective measure of trust in collaborative human-robot tasks is a challenging problem. The main issue is how to represent mathematically the concept of trust and measure the necessary parameters for computing a single score.

Our basic hypothesis is that "rational" trust behavior is reflected by the expected value of the decisions whether to allocate control to the robots on the basis of past robot behavior and the risk associated with autonomous robot control. The use of rational decision models for human–automation task allocation was proposed earlier using an expected-value analysis similar to a cost-benefit approach to improve diagnostic decisions. Sheridan and Parasuraman (2000) likewise propose several equations that can help improve decisions about allocating control to the human or the automation.

Using a rational decision model it is possible to establish a compound "goodness" score that collectively transforms the observed human task allocation decision behavior, risk and observed robot performance into an a relative expected loss score REL as shown in equation (1):

$$\textbf{(1)} \qquad \mathbf{REL} = \sum_{i=1}^{i=n} \mathbf{el_i / K}$$

Where $\mathbf{el_i}$ is the expected loss of robot autonomous control at trial **i;** $\mathbf{el_i}$ can be computed as

$$\textbf{(2)} \qquad \mathbf{el_i = P_i * C_i}$$

and

**n** is the number of trials

**K** is the number of operator overrides or intervention to take over control from the robot

The Relative Expected Loss can be written as

$$\textbf{(3)} \qquad \mathbf{REL} = \sum_{i=1}^{i=n} \mathbf{P_i * C_i / K}$$

We performed a less formal study to test the utility of our newly defined objective measure of trust. This study used one subject controlling the UGV over a wide variety of competency conditions in approximately the same scenario as before. Figure 6 shows the expected loss for a trial computed by Equation (3) on the x axis and the override score for that trial on the y axis.

The graphed results for individual trials separate clearly into three behavioral clusters, which we have termed 'over-trust,' 'under-trust' and 'proper-trust.' These are defined as:

1. An over-trusting operator overrides the robot a low number of times, even though the expected loss is high.
2. An under-trusting operator overrides the robot a large number of times, even though expected loss is low.
3. Proper-trusting of an operator is exhibited in the graph by the operator who overrides the robot a low number of times, when the expected loss is low and a large number of times when the expected loss is high.

These results were not subjected to statistical analysis and represent concepts that can be tested experimentally later.

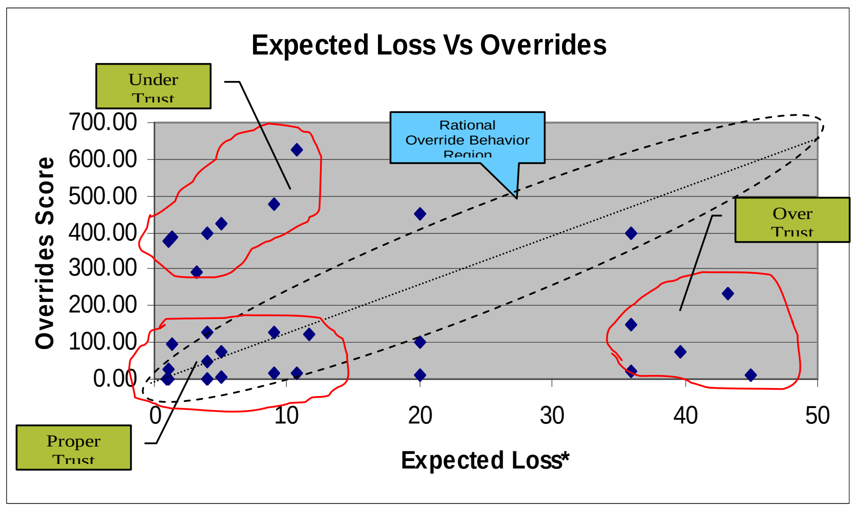


**Figure 6. Objective Trust Behavior**

The analytical methodology of Figure 6 can be used to diagnose specific operator trust behavior for the purpose of training and performance evaluation. The criteria for rational or "proper" behavior are represented by the oval region that contains the region in which expected loss should invoke an override. This region can be defined by exercising an expert with the system and obtaining the cluster of his behavior. This cluster can then be used as a standard for scoring operators.

The cluster location of a specific operator on the chart contains diagnostic value as to the characteristics of his behavior. As shown in Figure 6 some operators can be characterized as either over-trusting or under-trusting. To generate a measure of goodness for a specific operator, computational criteria can be created that measure his deviation from an expert. A quantitative measure of the "distance" between the operator relative expected loss and that of an expert provides direct indication for the type of feedback the operator needs during training in order to modify his trust behavior.

## SITUATIONAL AWARENESS

### Background and Measurement

The concept of situational awareness (SA) has been sub-divided into three levels:

1. Perception of Data
2. Comprehension of Meaning
3. Projection of the Near Future

Teams with high situation awareness have generally been shown to have improved performance (Cooke, Kriekel, & Helm, 2001).

In our study we used the SAGAT situational awareness assessment methodology developed initially for assessment of SA in air traffic control simulations (Endsley, 2000) and tailored it to our particular command and control scenario. Table 1 exemplifies the SAGAT questions asked during the scenario.

***Table 1* SAGAT Queries for MITPAS**

| | |
|---|---|
| . Enemy Detection | |
| a. | How many enemies does the UGV detect? |
| b. | Has the UGV targeted any enemies? |
| c. | Which enemy has been targeted by the UGV? |
| d. | Which enemy poses the largest threat? |
| II. Enemy Fire | |
| a. | Is the UGV currently taking fire from enemies? |
| b. | How many enemies are firing at the UGV? |
| c. | How many enemies have been killed by the UGV during the mission? |
| III. UGV Fire | |
| a. | Is the UGV firing? |
| b. | Is the UGV firing on the targeted enemy? |
| c. | Is the UGV firing on the nearest target? |
| d. | Will the UGV kill more enemies in the next minute? |

We hypothesized that as UGV firing competency decreases, situation awareness for the operator will also decrease. We further hypothesized that situation awareness would increase across trials. Finally, we hypothesized an interaction between fire competency and trial – i.e., that situation awareness would increase across trials, but least so for the low fire competency conditions.

### Findings

The results partially supported our hypotheses. The total SA score (all Levels) was not significantly different between conditions, but the total SA score may not be sensitive to subtle changes in SA as SA on queries can be independent of the other items (Endsley, 2000). This independence is particularly true for each of the three levels of SA, which are related to different operator information processing stages. For example, an operator may be able to detect a feature of the environment, but not comprehend how this element fits in and relates to the bigger operational picture. In addition, items have different response sets which may also reduce sensitivity.

Item-by-item analyses, however, did reveal significant results. As shown in Figure 7, Participants improved their situation awareness of the Level 1 SA query across trials. In the first trial, performance was about 60% and peaked around 80% in the fourth trial before dropping to 75% in the fifth trial.

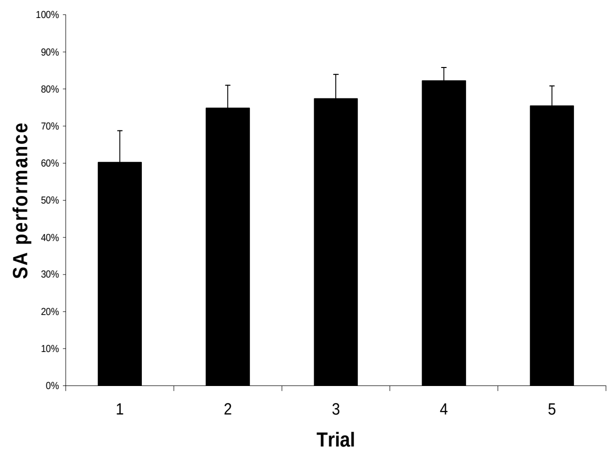


**Figure 7. Effects of trial on Level 1 SA**

As shown in Figure 8, Level 3 SA performance was relatively high (around 85%) in the high and medium UGV competency condition but was significantly lower (about 60%) in the low competency condition.

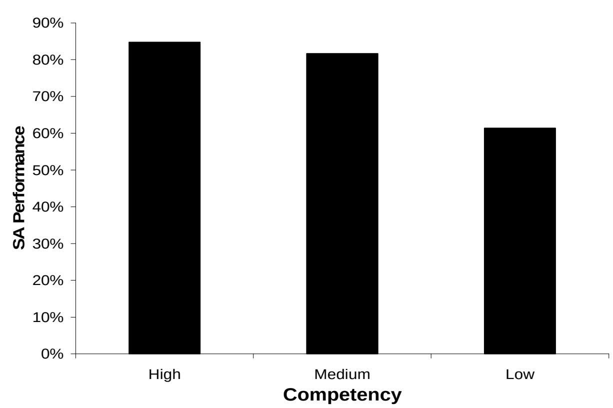


**Figure 8. Effects of competency on Level 3 SA**

The particular question in the Level 3 Situation Awareness (SA) query attempted to assess the participant's ability to project near-future courses of action based on current SA knowledge. It seems that in the low competency condition, participants were less likely to correctly project UGV behavior. This was likely due to the fact that UGV Fire behavior was less consistent (and therefore less predictable) in the low competency condition than in the other conditions. It is also consistent with our initial hypothesis.

Previous results in this study are consistent with these new findings. Mission performance was generally lower for the low fire competency condition compared to the other conditions, but improved considerably across trials. The results support the notion that situation awareness can be an indicator of team performance (Cooke et al., 2006).

Although we gained some support for our hypotheses, many SAGAT items did not reveal significant differences. A possible explanation for this finding may be that more measurements were needed to acquire a robust effect both by having more queries for each level of SA and number of measurements during the scenario (see also Cummings & Guerlain, 2007). In addition, each scenario lasted for only 10-15 minutes and this may have been too short to obtain an adequate, measurable SA picture. In future experiments, we intend to use an embedded and less obtrusive form of SAGAT to further increase the sensitivity of the measure.

## CONCLUSIONS

Numerous human factors studies have shown that successful human teams display particular characteristics of communication, coordination, and delegation that are merged together seamlessly to achieve the team goal. Many of these team processes can be measured and quantified for use in evaluating mission performance in different contexts or to evaluate training procedures (Cannon-Bowers & Salas, 1998).

As human-robot teams are increasingly developed and fielded, a similar approach to assessment of team performance must be undertaken. The MITPAS technology described in this paper represents a methodology for achieving this goal. Most previous work has focused either on robot performance alone (e.g., Albus, 2002), human performance alone (e.g., Wickens & Holland, 2000), or human-robot interaction (e.g., Parasuraman et al., 2005). We believe that MITPAS provides a unique and comprehensive methodology that incorporates all three of these elements.

Some specific findings from our initial studies that are relevant to both mixed initiative training and operations include the following.

• *Operators Will Compensate For Lack Of UGV Competence, But Only If The UGV Competence Level Is Clearly Identifiable.* Our results for Mission Time (the total time necessary to complete the mission) show that the operator learns to compensate partially for low competence of the UGV. Interestingly, there is little or no adjustment for medium competence; it seems that if the UGV is indeterminably competent, the operator has more difficulty adjusting his or her behavior to that of the semi-autonomous machine. Essentially the same results were seen for Kill Latency (the total time required for the human-robot team to destroy the target). Again, the operator seemingly adjusts better to a UGV with distinct (albeit poor) behavioral characteristics, even though his or her adjustment still does not bring the human-robot team to the best possible performance.

• *First Impressions Matter.* Our key measure here was Manual Distance, or the distance the operator covered with the UGV under manual control. This is also an objective measure of trust, because switching to manual control requires an override of the autonomous condition, implying a lack of trust that the UGV will accomplish its immediate task autonomously. The results indicated that for UGVs of either high or medium competence, the amount of Manual Distance is less if the operator first experiences a UGV of high or medium competence rather than one of low competence. Similarly, for UGVs of low competence, the amount of manual control is less if the operator first experiences a UGV of low competence. The conclusion is that it is best to train operators on a UGV with similar characteristics to the one he or she will be operating.

• *Familiarity Breeds Understanding.* Our results show that the overall subjective measure of Trust (simply asking the operator how well he or she trusts the UGV) correlates with the overall competency of the UGV, as might be expected. One result is counterintuitive, however, and should make us reassess what 'trust' actually means to Unmanned Vehicle operators. That is, our results indicated that subjective Trust increased with trial number for the low competence UGV (from a 1st trial level of 58 to a 5th trial level of 78, a 35% increase), while for the medium and high competence machines Trust remained relatively unchanged at levels of 85 and 90, respectively. It appears that trust concerns not just an expectation of *correct performance*, but also an expectation of *level of performance*. Once the operator knows that the UGV will make mistakes, he or she *trusts* it to make those mistakes. Once more, because the operator of the medium competency level UGV can not seem to get a good handle on what it will do, the subjective Trust level actually decreases slightly over 5 trials.

Further with respect to trust we have shown that objective measures of trust derived from MITPAS can be used to characterize operators as properly trusting, over-trusting or under-trusting. Such characterization, in addition to quantitative comparison with expert behavior, can have great diagnostic value for training feedback or for operational AAR.

Finally our initial trials with the SAGAT measure of situational awareness both helped validate our simulation environment in terms of hypothesis support, and revealed the likely need for improvements on this conventional approach in the context of human-robot mixed initiative team experimentation.

Overall, we believe that our initial results are indicative of the type of new insights into human-robot team behavior that can be gained by combining the measurement power of MITPAS with realistic simulations of tactical UV operations, such as that represented by our OneSAF-based experimental environment. Consequently, we intend to make the readily customized MITPAS available to other researchers and developers for use with their simulations, scenarios and special measures.

## ACKNOWLEDGEMENT

This research was supported by SBIR Phase II contract No. N61339-05-C-0003 funded by the U.S. Army RDECOM-STTC and administered in part by U.S. Army Research Institute, Orlando.

## REFERENCES

Adams, J. A., & Skubic, M. (2005). Introduction, special issue on human-robot interaction, *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, *35*, 433-437.

Albus, J.S. (2002). Metrics and performance: Measures for intelligent unmanned ground vehicles. *In MIS Proceedings*.

Barnes, M., Parasuraman, R. & Cosenzo, K. (In press). Adaptive automation for robotic military systems, Technical Report, NATO HFM, 2002

Cannon-Bowers, J.A., & Salas, E. (1998). Individual and team decision making under stress: theoretical underpinning. In J.A. Cannon-Bowers and E. Salas (Eds.), *Making Decisions Under Stress*. American Psychological Association, WDC

Cooke, N., Pringle, H. L. Pedersen, , H. K. & Connor, O. (Eds.). (2006). *Human Factors of Remotely Operated Vehicles. Advances in Human Performance and Cognitive Engineering*. Volume 7,pp. 285-297, Oxford, UK: Elsevier.

Cosenzo, K., Parasuraman, R., Novak, A. & Barnes, M., (2006). *Adaptive automation for robotic military systems*. ARL Technical Report, ARL-TR-3808

Cummings, M. L. & Guerlain, S. (2007). Developing operator capacity estimates for supervisory control of autonomous vehicles. *Human Factors, 49*(1), 1-15.

Endsley, M.R. (2003) "Direct measurement of Situation Awareness: Validity and use of SAGAT," In Endsley, M. R., Garland, D. J. (Eds.) Situation Awareness Analysis and Measurement. Mahwah, NJ: Lawrence Erlbaum Associates

Freedy, A., McDonough, J.G., Freedy, E.T., Jacobs, R., Thayer, S.M., and Weltman, G. (2004). A mixed initiative team performance assessment system (MITPAS) for use in training and operational environments. *SBIR Final Report, Perceptronics Solutions Contract No. N61339-04-C-0020*.

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*, 1243-1270.

Lee, J. D., & See, K. A. (2004) "Trust in automation: Designing for appropriate reliance." *Human Factors*, *46*(1), pp 50-80.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), pp 230-253.

Parasuraman, R., Galster, S., Squire, P., Furukawa, H., & Miller, C. (2005). A flexible delegation-type interface enhances system performance in human supervision of multiple robots: Empirical studies with RoboFlag. *IEEE Transactions on Systems, Man, and Cybernetics. Part A. Systems and Humans, 35,* 481-493.

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

Sheridan, T., & Parasuraman, R. (2006). "Human-automation interaction," *Reviews of Human Factors and Ergonomics*, 1, 89-129.

Wickens, C. D., & Holland, J. (2000). *Engineering psychology and human performance.*