

## A COMPREHENSIVE METHODOLOGY FOR ASSESSING HUMAN-ROBOT TEAM PERFORMANCE FOR USE IN TRAINING AND SIMULATION

E. De Visser, R. Parasuraman  
George Mason University  
Fairfax, Virginia

A. Freedy, E. Freedy, G. Weltman  
Perceptronics Solutions, Inc.  
Sherman Oaks, California

New methodologies and quantitative measurements for evaluating human-robot team performance must be developed to achieve effective coordination between teams of humans and unmanned vehicles. The Mixed Initiative Team Performance Assessment System (MITPAS) provides such a comprehensive measurement methodology. MITPAS consists of a methodology, tools and procedures 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 an initial experiment conducted to validate the measures and gain insight into the effect of robot competence on operator trust as well as on human-robot team performance.

### INTRODUCTION

#### UVs and Automation

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,” controlling 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. The goal of such 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).

Effective methodologies for evaluating human-robot team performance must be developed if the vision of a coordinated and effective team consisting of multiple human operators is to be realized. At the same time, understanding how a small team of operators can effectively control a large number of UVs with different types and capabilities requires analysis of issues related to levels of automation (Parasuraman, Sheridan, & Wickens, 2000), human-robot interfaces (Adams, 2005), and robot autonomy.

Current UVs can move and navigate autonomously, engage in goal-directed behaviors, and communicate with and provide feedback to human supervisors. Nevertheless, human supervision is essential both to specify high-level goals and to manage unexpected events. Despite ever-increasing advances in robot autonomy, the human-robot team must be a *mixed-initiative* system for the foreseeable future. In this paper, we describe the characteristics of mixed-initiative systems in the context of human-robot teaming and provide a comprehensive methodology for assessing the performance of such teams.

### PERFORMANCE METRICS

#### Performance Model

We have created a preliminary system performance model to identify the dimensions of performance which

contribute to effective outcomes of collaborative manned-unmanned tasks and in particular to formulate measures to evaluate processes that are unique to the collective team of humans and robots. We built our taxonomy on specific processes that can be decomposed into explicit behavioral objectives side-by-side with measures of effectiveness based on actual outcomes. We narrowed the performance measures to the simplest factor structure that adequately cover the dimension of teamwork as found in previous investigations (Cannon-Bowers & Salas, 1998). This structure can be decomposed into three distinct levels: operator performance, robot performance, and team performance.

#### Operator Performance

Objective measurement of operator performance include measures such as those based on robot tasking time, mission execution time or switching time can also be used (Parasuraman et al., 2005). In addition, subjective measurement can assess aspects of the control task such as situation awareness and workload.

#### Robot Performance

Performance of robots is typically measured by the efficiency of execution and navigation as well as by the robot's speed and reaction time to situational events. Such performance measures may include:

*Execution Efficiency.* The relative proportion of time that the robot is executing an operator instruction as opposed to the time it is waiting for direction as a function of total mission time.

*Navigational Efficiency.* The distance traveled by the robot from the start to the end of the mission as compared with the point-to-point distances summed along the pre-planned route.

#### Team Performance

Team performance that are currently being tested and evaluated in MITPAS include:

*Control Allocation.* Optimal allocation of tasks between the human and autonomous UV function. “Who should do the task and when?” Does the operator take control

where appropriate and let automation handle it when that is appropriate

*Human Robot Trust.* The level of trust at which the human delegates control to the UAVs. Is the operator over trusting or under trusting? Does he or she interrupt the UV unnecessarily due to lack of trust?

*Monitoring Feedback.* Observation and keeping track of UV's performance, including UV's mistakes. Recognizing good and bad performance by UVs; recognizing UV system conditions and operational status.

### VALIDATION EXPERIMENT

The present validation study was designed to replicate and expand on previous findings in automation research in order to validate our initial MITPAS metrics when used in conjunction with a realistic military scenario involving one unmanned ground vehicle (UGV) controlled by one operator. Our approach was to create an environment that introduced errors or failures in the robotic element of the team. This is because automation reliability is known to affect a number of dependent variables. Research has shown that an operator's trust in automation decreases as more failures are introduced, but that sometimes trust does not decrease if reliability is high initially (Fox & Boehm-Davis, 1998; Lee & See, 2004). Some studies investigating use of automated decision aids found that levels of trust will not be as high for initial low reliability as for initial high reliability (Fallon, 2005). If automation reliability is low enough, operators will likely engage in manual control of the task (Parasuraman & Riley, 1997). Mental workload increases when the operator decides to intervene and engage in manual control of robotic assets (Parasuraman et al., 2005). Furthermore, overall mission times increase when reliability decreases (Rovira et al., in press).

Robot reliability in this simulation was represented by UGV Firing Competency; its ability to shoot quickly and accurately at potential enemies. We hypothesized that as Firing Competency decreases, operator trust will also decrease. In addition, we expected that as UGV Firing Competency decreases, the operator's use of the available manual control mode would increase. Furthermore, we predicted that as UGV Firing Competency decreases, the workload of the operator will increase. Finally, we hypothesized that as the UGV Firing Competency decreases, mission and task times would increase.

### METHOD

#### Participants

Twelve young adults (4 females and 8 males) participated in a simulated battlefield mission. Most participants had several years of gaming experience and varied in age from 18 to 25. They were paid \$15 per hour.

#### Scenario and Task

The tactical scenario took place at a simulated location that had typical features of roads, forests and a small village or built-up area, shown in Figure 1.

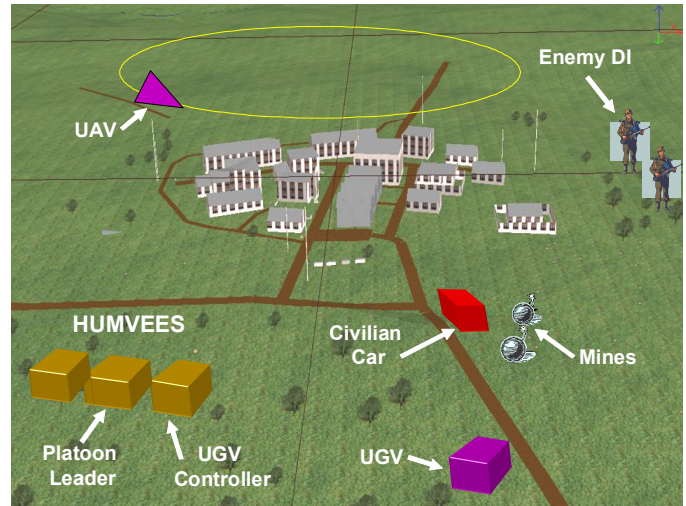


Figure 1. MITPAS Experimental Environment

The operator-UGV team was part of a reconnaissance platoon. Its mission was to insure that the area was safe by eliminating all surrounding enemies. To do this, the operator had to move the UGV to a checkpoint where it could commence targeting and firing on enemy forces until they were destroyed. Operators also had to monitor and evaluate the autonomous targeting and firing capabilities of the UGV and take over control of the vehicle if decrements in these autonomous processes would cause a mission delay or a complete failure; e.g., Firing Competency could be enhanced by moving the UGV closer to the enemy. 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.

#### Apparatus

The command and control configuration for our simulation environment and experimental study was as shown in Figure 2:

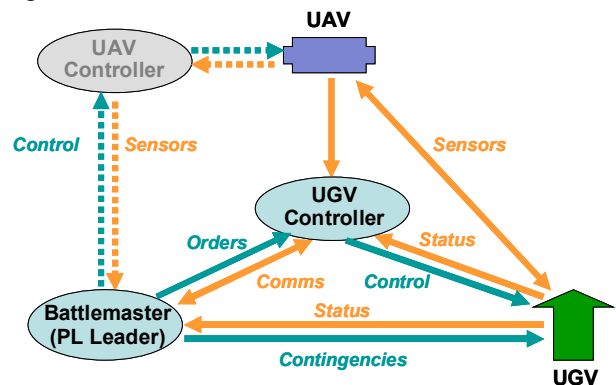


Figure 2. Command and Control Configuration

The Battlemaster, who also played the platoon leader, was in charge of the experimental procedures, the progress of the scenario, and communication with the Unmanned Ground Vehicle controller, who was the experimental participant.

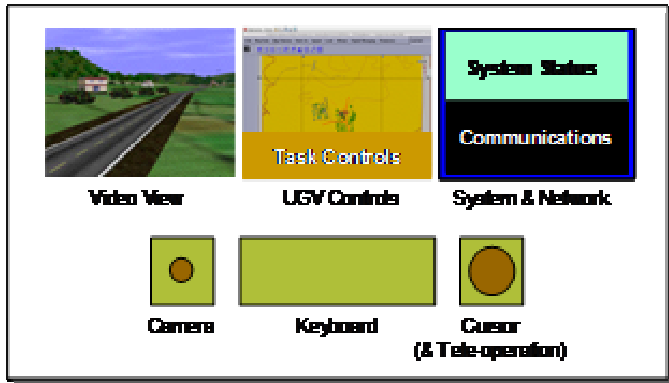


Figure 3. UGV Controller Station

The UGV controller station shown in Figure 3 was comprised of three 19 inch monitors, a keyboard, track-ball mouse, and joystick. The tactical situation map, generated by the military simulation OneSAF 2.5 on the center monitor, showed the UGV's geographic position and also gave an overview of the entire tactical situation. A 3D representation of this environment was simulated by a MAK Stealth 5.4 on

the left screen and showed live action from the viewpoint of a camera mounted on the UGV. The right monitor display gave real-time information about UGV status, including its direction, current target, surrounding friendly and enemy units, supply levels, and any vehicle malfunctions. Communications between the operator and the Battlemaster were facilitated through instant-message capabilities.

The UGV control interface enabled the operator to assign the UGV specific automation tasks from an abbreviated OneSAF menu. The operator could also tele-operate the UGV using the joystick in combination with gas and brake pedals.

**Procedure**

The experiment was a 3x5x6 mixed factorial design with UGV Firing Competency (high, medium, low) and Trial (1 through 5) as the two within-subjects factors and the six possible orders of UGV Firing Competency as a between-subjects factor. UGV target malfunctions varied randomly based on algorithms of UGV behavior in the ONESAF simulator. Variability caused by these target malfunctions was eliminated in the data analysis because these malfunctions

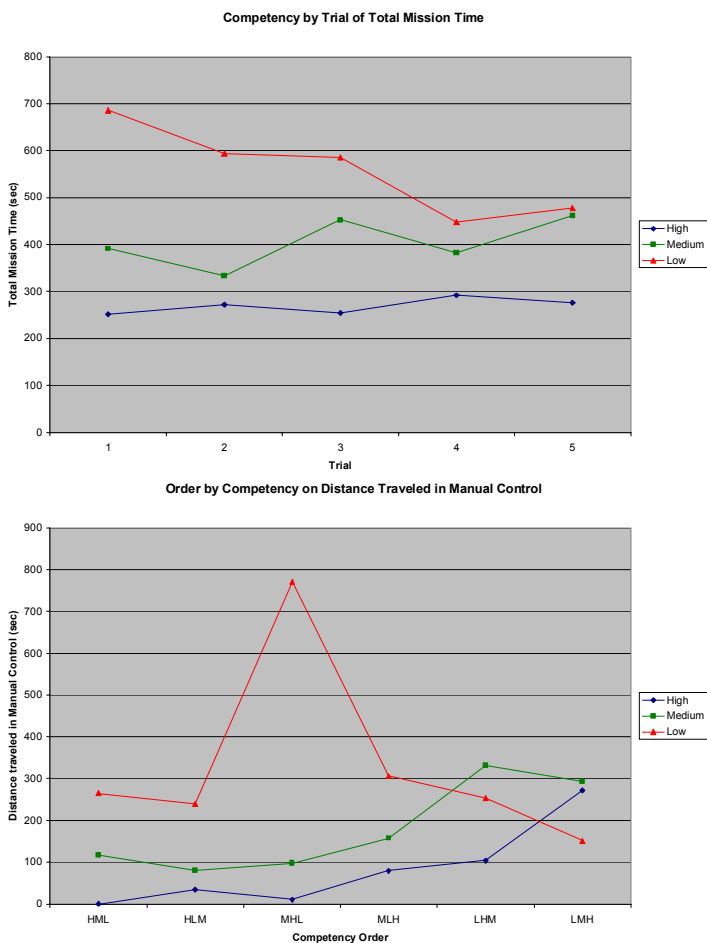
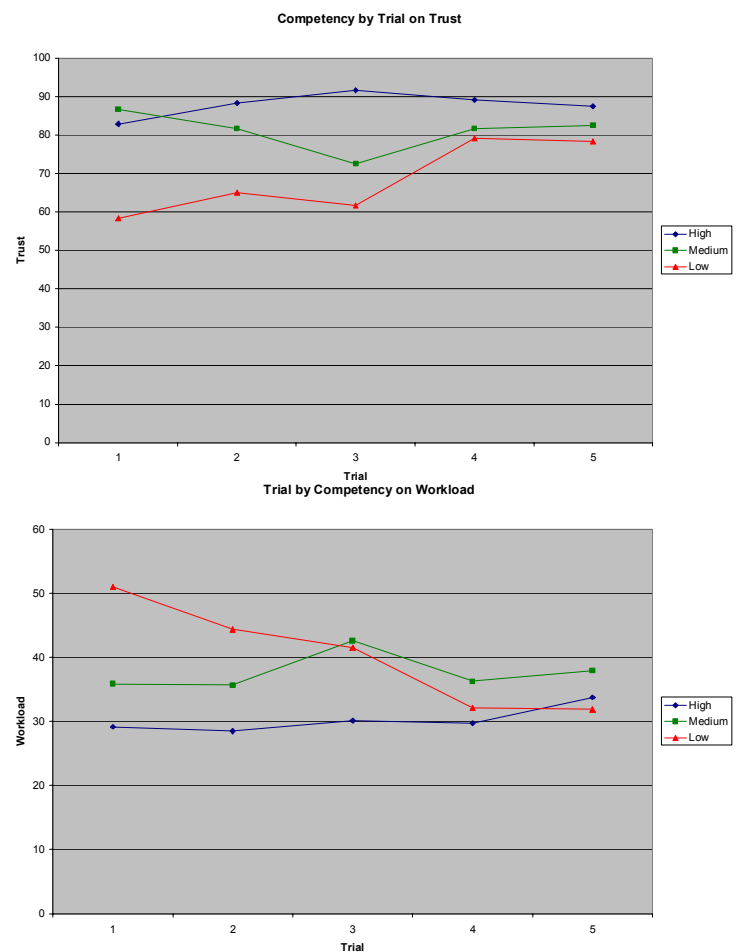


Figure 4. Results are shown for (a) Mission Time vs. Trial (top left), (b) Distance Traveled in Manual Control vs. Competency Order (bottom left), (c) Trust vs. Trial (top right) and (d) Workload vs. Trial (bottom right)



were not controllable by the researchers. The dependent variables of interest included total mission time, time latencies between enemies detected and killed, control allocation, human robot trust, situation awareness, and workload.

Participants first received about 2.5 hours of training on proper usage of the MITPAS system from a comprehensive training manual and supplemental instruction as needed. Participants then completed 5 trials of each level of UGV Firing Competency 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. After completion of each trial, participants filled out the NASA TLX for subjective workload and an adapted version of a common trust and self-confidence measure used in earlier automation research (Lee & Moray, 1992). Situation awareness was measured by SAGAT queries specifically tailored to this scenario (Endsley, 2000). The mission was frozen during each trial at about half of the expected overall mission time for each reliability condition based on the results of an earlier pilot study.

## RESULTS

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.

### Total Mission Time

There was a significant interaction between Competency and Trial for total mission time,  $F(1,8) = 3.83$ ,  $p < 0.05$ . In the low Competency condition mission times decreased across trials while mission times in both the high and medium Competency condition stayed constant, see Figure 4(a). In addition, overall mission times increased as Competency decreased. Similarly, Competency was shown to have a significant interaction with Trial on the average time to kill an enemy,  $F(1,8) = 3.68$ ,  $p < 0.05$ .

### Manual Control

An ANOVA on the total distance traveled in manual control showed a significant interaction between Competency Order and Competency,  $F(1,10) = 8.68$ ,  $p < 0.05$ . Total travel distance in manual control was higher for orders that started trials with low Competency for both high and medium Competency conditions than for orders that began trials with high Competency, see Figure 4(b).

### Subjective Measures

A significant interaction between Firing Competency and Trial was found on subjective trust,  $F(1,8) = 4.68$ ,  $p < 0.001$ . In the low Firing Competency condition, trust increased across trials. In the high and medium Firing Competency condition trust remained at the same mean, see Figure 4 (c).

There was also a significant interaction between Firing Competency and Trial on workload,  $F(1,8) = 4.90$ ,  $p < 0.001$ . Workload decreased in the low Firing Competency condition across trials, but did not differ in the high and medium Competency conditions, see Figure 4(d).

## DISCUSSION

Consistent with our hypothesis, overall mission time increased as UV competency decreased. In the first trial of the low competency condition, the average mission time was 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 progressed, overall mission time decreased 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, increasing its firing accuracy and resulting in a shorter overall mission time. This effect is also confirmed by the decreasing average time it took to kill an enemy in the low competency condition.

The subjective workload and trust measures appear to support the observed decrease in mission times for low competency levels across trials. Initially high workload in the low competency condition decreases as trials progress, reaching the levels of workload in the medium and high competency conditions. This effect is consistent with the idea that participants become more capable of handling the unreliable UV in manual control. Inconsistent with our initial hypothesis, trust increased across trials in the low competency condition. This finding could be explained by operators moving the UV closer to the enemy and thus making it more accurate, increasing trust in the UV's capability to eliminate enemies. In addition, it may reflect a nuance in the operators' definition of trust; that is, as the operator becomes more familiar with the unreliable robot, he or she learns to anticipate its poor performance and also his or her ability to compensate -- thus increasing the reported trust level.

The effects of Firing Competency and Trial on trust, workload, and overall mission times provide converging evidence that the operator learned 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. Consequently, trust increased because the UV becomes more accurate when moved closer to the enemy and workload decreased as operators become more skilled in controlling the robotic asset. 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. This indicates that first impressions do make a difference; that is, operators may be

less patient and less trusting of an UV after initially experiencing low robot competency, making them insensitive to subsequent improvements in competency. This effect could have particularly damaging repercussions for human-robot interaction when applied to a real-world setting in which automation starts out with low reliability and generally becomes better over time, for example through adaptive mechanisms.

## CONCLUSIONS

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). MITPAS provides a unique and comprehensive methodology that incorporates all three of these elements.

We believe 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. We intend to make the readily customized MITPAS available to other researchers and developers for use with their simulations, scenarios and special measures.

## ACKNOWLEDGEMENTS

The authors thank M. Kalphat, D. Palmer and N. Coyeman for their help with the overall MITPAS development and evaluation project and also thank D. Boehm-Davis, P. McKnight, V. Kapur for their useful comments and advice on the validation experiment and this paper

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, FL.

## REFERENCES

Adams, J. A. (2005). *Human-robot interaction design: Understanding user needs and requirements*. Proceedings Human Factors and Ergonomics Society 49<sup>th</sup> Annual Meeting.

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.

Clough, B.T. (2002). Metrics, schmetrics! How the heck do you determine a UAV's autonomy anyway? *AFB in MIS Proceedings*. Air Force Research Lab, Wright Patterson.

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

Endsley, M.R. 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.

Fallon, C. K. (2005). *Improving user trust with a likelihood alarm display*. Paper presented at the Proceedings of the 1st International Conference on Augmented Cognition, Las Vegas, NV.

Fong, T., Kaber, D., Lewis, M., Scholtz, J. Shultz, A., and Steinfeld, A. (2004). A common metrics for human-robot interaction. *IEEE International Conference on Intelligent Robots and Systems*, Sendai, Japan.

Fong, T., Thorpe, C. and Baur, C. (2002). Collaboration, dialogue and human robot interaction. *In proceedings of the 10<sup>th</sup> International Symposium on Robotic Research*, Spinger Verlag.

Fox, J. E. & Boehm-Davis, D. A. (1998). Effects of age and congestion information accuracy of advanced traveler information systems on user trust and compliance. *Transportation Research Record 1621 (Safety and Human Performance)*. Washington, D.C.: National Academy Press, 43-49.

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 Phase I Final Report, Perceptronics Solutions Contract No. N61339-04-C-0020*.

Kalphat, H. M., and Stahl, J. (2002). STRICOM's advanced robotics simulation STO: The army solution to robotics M&S. *In proceedings of the eleventh conference on computer generated forces & behavioral representation*, Orlando, FL.

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

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.

Rovira, E., McGarry, K., Parasuraman, R. (in press). Effects of imperfect automation on decision-making in a simulated command and control task. *Human Factors*.

