

Implementation of a Situation Awareness Assessment Tool for Evaluation of Human-Robot Interfaces

Jean Scholtz, *Member IEEE*, Brian Antonishek, Jeff Young

Abstract— In this paper we outline a methodology for evaluating the situation awareness (SA) provided by a supervisory interface for an autonomous on-road vehicle. Our goal is to be able to use the evaluations to compare interface designs with respect to how well each facilitates the users' acquisition of situation awareness. We used Endsely's Situation Awareness Global Assessment Technique (SAGAT) [8] and developed scenarios and assessment questions appropriate for supervisors of autonomous on-road driving vehicles. We describe the results of two experiments used to refine our SA assessment implementation. In a third experiment we applied the refined implementation to a graphical user interface we developed to test the sensitivity of our SAGAT implementation. We discuss the results of this experiment and implications for applying the SAGAT methodology to supervisory user interfaces for autonomous vehicles.

Index Terms— human-robot interaction, evaluation, situation awareness, supervisory user interface

I. INTRODUCTION

User interfaces for robots fill several roles: they provide a mechanism for a user to give input or commands to a robot; they provide information to the user about the status of the robot; and they provide information about the remote environment in which the robot is operating. As robots move out of the research laboratories and into operational environments, these interfaces will be used by users who are not robotics experts. In order to develop usable interfaces, we need to find techniques for evaluating human-robot interfaces.

Usability engineering for desktop computers has been developing evaluation techniques and metrics for the past twenty years. ISO 9241[1] defines usability as the effectiveness, efficiency, and satisfaction with which specified users achieve specified goals in particular environments. The question is whether current methods of usability engineering

and human-computer interaction (HCI) are adequate to assess interfaces designed for human-robot interaction (HRI)?

There are some key differences between human-computer interaction and human-robot interaction. One difference we are interested in is the number of individuals who might interact with a robot and the roles those individuals might assume [2,3]. Users can interact with robots as operators, supervisors, peers, bystanders, or mechanics and programmers. Each of these roles needs different information and different interactions. It is possible that one person could assume more than one role. This would necessitate a user interface that presented the required information for the roles assumed and supported the required interactions or controls. It is also possible that a number of different users could interact with the robot simultaneously, each assuming different roles.

Another key difference is that many times, the user and the robot are not co-located. Therefore, part of the information the user may need is an understanding of the remote environment. This is particularly true for the roles of the operator and the supervisor, as individuals in these roles are most likely to be remote. Designers of user interfaces for teleoperation use the term "telepresence" to refer to providing the robot operator with the feeling of "being in the environment." Teleoperation user interfaces rely heavily on video from cameras placed on the robot. It is difficult for operators to navigate, even in familiar environments, when cues to distance and location are degraded by a narrow field of view and abrupt changes in field of view. Cameras are often close to the ground which provides a perspective most of us are not used to. Even familiar objects can be difficult to recognize under such conditions. In addition, the environment often has undesirable conditions, such as poor lighting or smoke, that impair visual perception. Communication issues can also result in degraded video being transmitted to the operator.

We are interested in evaluating interfaces for the supervisory role. Our current work focuses on the domain of autonomous vehicles in on-road driving situations. The supervisor would be responsible for monitoring a number of vehicles and either intervening when a vehicle encounters a problem or handing the vehicle off to an operator if the supervisor does not currently have the resources to handle the intervention. We are able to use traditional usability testing methods to assess the interactions necessary for intervention. However, we currently lack a way to assess the presentation

Manuscript received August 1, 2004. This work was supported in part by the DARPA MARS program. Jean Scholtz is with the National Institute of Standards and Technology, Gaithersburg, MD 20899 USA (phone 301-975-2520; fax: 301-975-5287; e-mail: jean.scholtz@nist.gov)

Brian Antonishek is with the National Institute of Standards and Technology, Gaithersburg, MD 20899 USA. (e-mail: brian.antonishek@nist.gov)

Jeff Young was with the National Institute of Standards and Technology. He is now with Resource Consultants, Inc. Vienna, VA. 22180 USA.

of the information the supervisor uses to determine if an intervention is needed or is likely to be needed soon. That is, we need to assess the situation awareness provided by the user interface.

II. SITUATION AWARENESS

Endsley [4] defines three levels of situation awareness: perception, comprehension, and projection. Perception is the basic level of situation awareness (SA level 1). This level of awareness is achieved if operators are able to perceive in the user interface the information that is needed to do their job. The next level is comprehension (SA level 2). Not only must the information be perceived, it must be combined with other information and interpreted correctly. The third level (SA level 3) is projection or the ability to predict what will happen next based on the current situation. As situations are dynamic, time is critical to situation awareness as well. User interfaces need to be designed to facilitate the continuous acquisition of SA.

III. EVALUATING SITUATION AWARENESS

Operator interfaces for control of semi-autonomous systems, such as aircraft, power plants, and manufacturing systems, have been assessed for situation awareness. Performance based, knowledge-based, subjective ratings, or direct assessment methodologies have been used to evaluate operators' SA.

Performance methods look at the outcome [5]. Did the system (consisting of the user and the robot) perform correctly given the situation? While we are interested in the end result, there are problems with attributing incorrect behavior solely to a lack of situation awareness or even attributing correct behavior to good situation awareness. Users are not perfect; even given good situation awareness, they can make inappropriate decisions and issue commands that are inappropriate. Factors in the environment or problems with the robot can prevent an appropriately selected plan from being completed. Users with a lack of situation awareness may also be able to select an appropriate behavior.

Knowledge-based methods are used in experimental conditions to isolate particular components and assess them individually. Knowledge-based methods are better at uncovering declarative information than procedural information. Verbalization methods, such as think-aloud and talk-aloud protocols [6] are also used to discover what information users are relying on when making their decisions.

Subjective measures [7] ask users to assign a numerical value to their situation awareness at any given time. While this gives an indication of the level of awareness, it fails to help developers understand what information is missing. This method, however, can be used in an operational environment.

IV. DEVELOPMENT OF A METHOD TO MEASURE SITUATION AWARENESS IN HRI

Our work is modeled after the Situational Awareness Global Assessment Technique (SAGAT) developed by Endsley [8]. This evaluation methodology uses expert knowledge to develop questions that assess the users' awareness of a particular situation. The methodology uses a simulation. The user is stopped during the simulation and given a quick series of questions to answer. These questions assess the three levels of situational awareness. After answering these questions, the users are returned to the simulation.

We felt that the SAGAT methodology was appropriate to implement for several reasons. First, we wanted to use the technique in formative evaluations and separately from the assessment of the user interaction. We did not want to have a system fully integrated with robots for such early evaluations. We wanted to use the evaluations to compare interface designs with respect to how well each facilitates the acquisition of situation awareness by users.

In the following sections we describe the implementation of the SAGAT methodology, the two experiments we used to refine the implementation, and a third experiment we used to assess the sensitivity of the implementation.

V. DEVELOPING THE HUMAN-ROBOT INTERFACE

We selected on-road driving as the domain. While we envision that the final HRI will support multiple vehicles, we started the implementation with only one vehicle as we felt it was important to develop our SA questions appropriately for one vehicle. Then we would move to the case for multiple vehicles. However, we did design our user interface so that the features were scalable to multiple vehicles.

Fig. 1 shows the initial user interface we developed. The assumption is that sensors on the vehicle can provide information such as the number of cars around the vehicle, current traffic controls, obstacles perceived in the roadway, and pedestrians nearby if the vehicle is in an urban environment. Also, the vehicle would provide information such as speed and amount of fuel left. We used a map background for the HRI. We used three blocks of text to display the detailed information about the vehicle, the environment, and the route. The detailed information about the vehicle and the environment can be closed if the user wishes. All the windows can be moved around to suit the user's preference. For multiple vehicles, the detailed information could be designed as tabs. The supervisor could select the correct tab for the vehicle she wished to see, or a particular tab could be displayed automatically if there was an issue with that vehicle.

We used two superimposed symbols to show the status of the vehicle and the surrounding environment. The outer symbol represented the status of the environment surrounding the vehicle while the inner symbol represented the status of the vehicle. We used both symbols and color to reflect a changing status so we did not make assumptions about the color perception of the supervisors. Fig. 2 shows the icons used in the HRI to display the various conditions of the

vehicle and the environment. In our experiment, the conditions for changing the status were preset but in an operational environment, the supervisor would have control over this.

traffic, etc. The GPS data points were stored once/second and were presented at the same rate in the simulation.

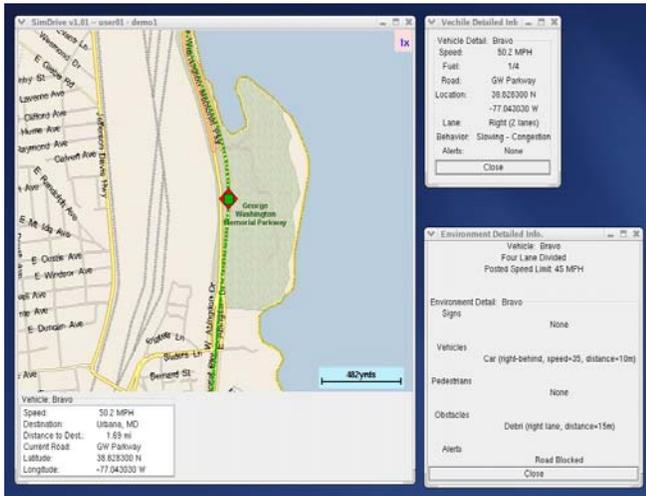


Figure 1. The initial supervisory HRI used in the experiment

STATUS ICONS	Vehicle Normal	Vehicle Caution	Vehicle Trouble
Environment Normal			
Environment Caution			
Environment Trouble			

Figure 2. The icons used to reflect the status of the vehicle and the surrounding environment

VI. DEVELOPING THE SIMULATION AND THE SCENARIOS

In order to conduct our experiments, it was necessary to implement a simulation. The 'SimDrive' simulator program was written starting with 'GPS-Drive' [9] as a base to utilize its routines for displaying maps and for working with GPS data points. GPS-Drive is a GPS navigation system which runs on Linux and is written in C using the GTK+ graphics library. GPS-Drive is written under GPL (GNU General Public License) [10].

We implemented routines to view simulated driving scenarios and created a file editor to add details (such as traffic) to the scenarios. For our experiments, we created simulation data files using the file editor. We used a handheld GPS unit to collect position data for a number of routes within a 40 mile radius of our laboratory. We collected this data on different roads at different speeds. Then we selected the segments we wanted to use in the experiment and used the file editor to input different conditions, such as the amount of

Table 1: Scenarios used in our experiments

#	Speed limit	# Lanes	Vehicle in Lane	Risks at freeze point during Experiment 3	Conditions encountered during scenario
1	45	4	right	Red light, slow speed	Red light; another car
2	55	6	right	Excess speed	Speed limit exceeded entire time
3	45	4	left	Congestion, slow speed, 2 cars in other lane	Speed limit exceeded; two cars encountered
4	45	4	right	Debris blocking lane; car beside, slow speed	Cars encountered
5	55	6	right	Car in front	Speed limit exceeded
6	55	4	right	Car to left; car in front	nothing
7	55	4	right	Exit in a mile; slow speed	Low fuel; exit at 2 miles
8	45	4	right	Lane blocked by debris; debris on left roadside; car beside, slow speed, congestion	Car passing; traffic stopping
9	45	4	left	Debris in front; car behind, slow speed, congestion	Pedestrian and two cars; vehicle shifts lanes;
10	45, 35, 45	4	left	Pedestrian on roadside	Speed limits change; right turn made

We developed 10 scenarios that represented different types of driving conditions and different hazards. We used scenarios from freeway driving and from highway driving. We have not yet incorporated scenarios for rural driving, driving in suburban neighborhoods, or driving in congested city streets. Rural driving poses fewer problems, and driving in suburban neighborhoods and congested city streets present more many more problems for autonomous driving vehicles.

In the two types of driving scenarios we constructed, drivers need to be aware of traffic in parallel lanes, traffic ahead and behind their vehicles, exits and on-ramps, speed limits, debris in roadways, traffic controls on highways, and pedestrians at intersections. Drivers also need to have knowledge of the condition of their vehicle, such as fuel remaining or engine lights that are on. Route information is also important. Is an exit that should be taken approaching? How close is the destination?

Table 1 describes the scenarios that we used for our experiments. We have a database that holds the “ground truth” about the scenario. That is, we have data points and speed data from the GPS unit. Using the simulation editor we can insert the different environmental conditions, such as traffic, and the destination information compatible with the GPS data points.

VII. SA ASSESSMENT QUESTIONS

Following the SAGAT methodology, we used our ‘SimDrive’ simulator to show a driving sequence to subjects. At a certain point in time the simulator is frozen, the screen goes blank, and the subject is asked to turn to another computer and respond to a set of situation awareness questions based on the three levels of SA as defined by Endsely.

The first step in developing questions for the three levels of SA is to draw on experts in that particular domain. As we were working in the on-road driving domain, we used a computerized driving training program as the source for our questions (Driver-ZED¹). We used direct questions about the vehicle status, the environment, and the status of the mission as SA level 1. These were given as multiple choice questions. There is only one correct answer for each of the three questions about status. For SA level 2, we asked the subjects to determine the potential risks in this situation. They were asked to select these risks from a list that was displayed. The lists of risks to select from were always the same but the actual risks were different for each scenario. For SA level 3 probes, we asked subjects to respond “yes” or “no” to questions such as “could you turn left right now?” There was only one SA level 3 question for each situation which was based on the situation at the time of the freeze.

VIII. EXPERIMENT ONE

The objective of the initial experiment was to determine how well our SAGAT implementation worked.

A. Subjects

Ten subjects, all of whom held valid drivers’ licenses, were recruited from our co-workers. Nine males and one female participated.

The screenshot shows a user interface for a simulation experiment. It is divided into several sections:

- Vehicle situation is:** Radio buttons for Normal, Cautionary, Dangerous, and I don't know.
- Environment situation is:** Radio buttons for Normal, Cautionary, Dangerous, and I don't know.
- Distance to destination:** Radio buttons for <1 mile, >1 and <5 miles, between 5 and 10 miles, greater than 10 miles, and I don't know.
- Identify Current Risks:** A list of checkboxes for: People nearby, Light turning red, Vehicle behind, Vehicle ahead, Vehicle on the left, Vehicle on the right, Speed too fast, Speed too slow, Obstacle on the roadway, Parked vehicles on the roadside, Exit approaching, and Merge ahead.
- Demo_Q1: Is it safe to accelerate?** Radio buttons for Yes, No, and I don't know.
- A **Submit** button at the bottom.

Annotations with arrows point to these sections:

- SA Level 1 Questions:** Points to the Vehicle and Environment sections.
- SA level 2 Questions:** Points to the Risks section.
- SA level 3 Question:** Points to the Demo_Q1 section.

Figure 3. Data collection user interface for experiment one

B. Experimental Design

Subjects were told that they were monitoring a remote autonomous driving vehicle and that as supervisors they needed to know if the vehicle was having trouble at any given time. They were told that this experiment involved only a simulation of this vehicle. They were shown the user interface and the meaning of the various icons was explained. They were told that at a certain point in time the simulation would stop and they would be asked questions about the condition of the vehicle, the route, and the traffic conditions. Subjects were shown how to display the windows for the condition of the vehicle and the environment. All subjects choose to have these on the screen during the experiment. They were allowed to arrange the windows however they wished. We developed three demonstration scenarios that we presented to subjects to train them in the procedure for the experiment. Then we presented the 10 driving scenarios, each lasting 1-2 minutes. After each scenario, both in the training and in the actual experiment, the subjects were given the situation awareness questionnaire screen for the scenario they had just monitored. The scenarios were presented in randomized order to each of the subjects to eliminate order effects. At the end of the 10 scenarios, we gave the subjects a questionnaire asking them to rate the difficulty of assessing the condition of the vehicle, the environment, and the route.

IX. RESULTS

Tables 2, 3, and 4 show the results for the three levels of SA that we assessed in this experiment. For SA level 1, there were three points for each scenario as we asked one question about the status of the vehicle, one about the status of the environment, and one about the status of the route. The results for SA level 1 were encouraging.

¹ Use of this product does not constitute endorsement by the National Institute of Standards and Technology

Table 2. The percent of incorrect responses for the 3 categories of SA level 1 for experiment one

Scenario	Vehicle	Environment	Route
1			10
2			10
3		10	10
4		10	20
5			30
6			20
7	30	10	
8		20	10
9		40	10
10			20

Table 2 shows the incorrect responses for the 3 categories of SA level 1 we were testing. If the table entry is blank, that indicates 100% correct responses. Subjects had little trouble with identifying the condition of the vehicle except for scenario 7. In this scenario, the vehicle's condition went red at the very end, so subjects may not have had enough time to recognize this. The environment and route responses were more problematic. One potential issue with the route is that the information was presented textually and in a different location in the user interface than the vehicle and environment information. All the scenarios where subjects incorrectly answered the environmental SA question had different conditions for the vehicle and the environment. However, this was also true for scenario two. One possible explanation is that subjects relied more heavily on the textual description than on the icon indicators.

Table 3. SA level 2 responses for experiment one

Scenario	# of risk indicators	omissions– Mean(SD)	additions– Mean(SD)
1	1	.20 (.42)	.10 (.32)
2	1	.20 (.42)	.00 (.00)
3	2	.30 (.48)	.00 (.00)
4	2	.40 (.51)	.10 (.32)
5	2	.70 (.67)	.10 (.32)
6	3	.50 (.53)	.00 (.00)
7	1	.00 (.00)	.10 (.32)
8	1	.10 (.32)	.30 (.48)
9	3	1.30 (.67)	.00 (.00)
10	1	.10 (.32)	.00 (.00)

For SA level 2 each scenario had a different number of risks. There was not merely a right or wrong answer. Subjects could omit risks and add risks. That is, they might fail to mark a risk that we identified or they might add a risk that was not actually present in the scenario.

Table 3 shows the mean number of omissions and additions for each of the 10 scenarios. Omissions were the primary source of error as opposed to additions. As the number of risks increased so did the number of omissions. This could possibly be attributed to the ability of subjects to recall all the risks. We were also concerned about the descriptions of the risks and how those were interpreted by the subjects. For example, scenarios 5 and 9 were more problematic than the

others. One of these scenarios had an obstacle in the road. Both scenarios described vehicles in relationship to the vehicle being monitored. The text description for scenario 5 was:

“truck (right-ahead, speed = 55, distance = 10 m)”

This was meant to be interpreted as a vehicle that was on the right of the autonomous vehicle and ahead of it by 10 m.

Scenario 9 had the following description of the Obstacle:

“debris (right lane, distance = 15 m)”

The subjects needed to also note that the vehicle description showed the autonomous vehicle in the right of 2 lanes. Therefore, the debris was in front of the vehicle.

Table 4. The results for SA level 3 from experiment one

Scenario	% Correct	% Didn't know
1	70	20
2	80	10
3	80	
4	90	
5	50	
6	80	10
7	100	
8	50	
9	40	10
10	40	10

Table 4 shows the results for SA level 3. We gave subjects the option of selecting “I don't know” to eliminate guessing. As the table shows, 6 answers in total were “I don't know”. Scenarios 5, 8, 9 and 10 were the most problematic. The SA level 3 question for scenario 9 was:

“is it safe to break suddenly to avoid an obstacle?”

The answer was “no” as there was a car directly behind the autonomous vehicle. Not surprisingly, the same scenarios that caused problems with SA level 2 also caused problems with SA level 3. If subjects have difficulty recognizing the risks present for level 2, they will be unable to correctly make predictions about safe courses of actions to take.

Table 5. Post-experiment questionnaire for experiment one

Question	Mean rating (1 – extremely easy, 7 – extremely difficult) (SD)
Determining overall condition of vehicle	2.7 (1.2)
Determining overall condition of environment	2.7 (1.2)
Determining overall condition of route	2.4 (1.1)
Awareness of situation (1 – always aware, 7 never aware)	3.0 (0.8)
How many cars do you think you could monitor with this UI?	1.3 (ranged from 1 to 3) (0.7)

Table 5 shows the results of the post-experiment questionnaire. In general, subjects felt that it was relatively easy to determine the condition of the vehicle, environment, and route (SA level 1). However, they did not feel that they would be able to monitor many more vehicles using this user interface.

X. EXPERIMENT TWO

After studying the results of experiment one, we decided that we needed to revise our assessment questions for SA level 2 and level 3. First, we needed to revise the way that we asked subjects to indicate risks for SA level 2 as results from experiment one seemed to indicate that this was confusing. We wanted to make the assessment more precise to indicate for example, how subjects were to note that a car was to the left and ahead of the vehicle. In the first experiment we expected subjects to check multiple boxes to indicate a situation such as a car is ahead in the left lane. We were not sure that our subjects always remembered this even though it was explained in the introduction and training. We redesigned the data collection portion for SA level 2 to clarify this. For SA level 3, we gave subjects three choices: yes, no, I don't know. However, we were not sure if their yes or no was based on the correct awareness of the situation, so we added an explanation box for subjects to explain why they answered as they did. We did not change the SA level 1 questions, the scenarios or the user interface design.

A. Subjects

We recruited 3 subjects from co-workers. All subjects held valid drivers' licenses. We used only a few subjects as we regarded this experiment as more of a pilot to test the newly designed collection user interface prior to running the experiment again. Two of the subjects were males and one subject was female.

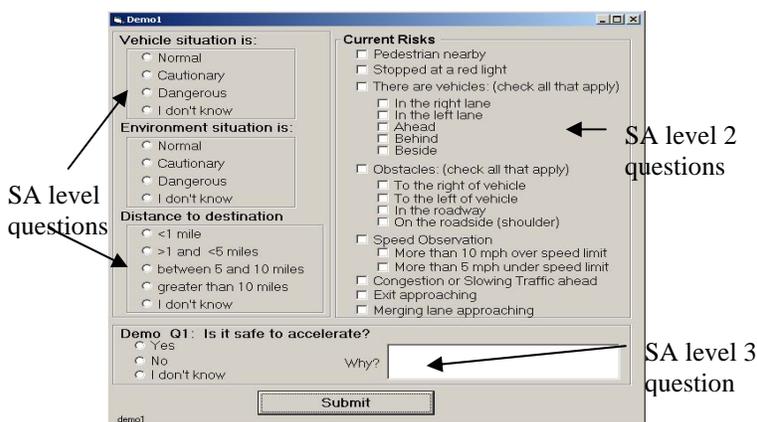


Figure 4. Data collection UI for experiment two and three

B. Experimental Design

We used the same explanations and training as in experiment one. We again presented the 10 scenarios in randomized order so that all subjects saw them in a different order. However, the questionnaire assessment for level 2 SA was changed to be more precise. For the SA level 3 question, we

added an explanation block of text. We wanted to use this to determine the subjects' interpretations of the information. The new situation assessment questionnaire is shown in Fig. 4.

XI. RESULTS

We made no changes to the questions or data for SA level 1 so we expected to see very similar results to those in experiment one if our SA methodology is repeatable. Table 6 shows the results for SA level 1 questions.

Table 6. The percent of incorrect responses for the 3 categories of SA level 1 for experiment two

Scenario	Vehicle	Environment	Route
1			
2			67
3			33
4		33	33
5			33
6		33	
7	33	33	
8	33	33	
9		33	
10			

Note that in Table 6, a blank indicates that all of the responses were correct. As with experiment one, the subjects had less trouble with the SA level of the vehicle than with the environment or the route. In experiment one, scenarios 3, 4, 7, 8 and 9 received incorrect answers for environment. In this experiment, scenarios 4, 6, 7, 8 and 9 received incorrect answers. Route information seemed less troubling for these subjects, however. Given the small number of subjects we used in this experiment, we hesitate to make assumptions about the repeatability of the SA level 1 portion of the experiment, but we feel that results look promising.

The big change we were investigating was the way we were eliciting answers for SA level 2. While we made the specification of the risks more precise in the way we asked subjects to indicate them, we also greatly increased the number of boxes that subjects had to check to indicate the situation.

Table 7. SA level 2 responses for experiment two

Scenario	# of risk indicators	omissions– Mean(SD)	additions– Mean(SD)
1	2	.00 (.00)	.00 (.00)
2	1	.33 (.58)	.00 (.00)
3	6	1.66 (1.15)	.00 (.00)
4	7	5.33 (1.53)	.33 (.58)
5	3	1.33 (1.53)	.00 (.00)
6	5	1.33 (1.15)	.00 (.00)
7	2	1.33 (.58)	.00 (.00)
8	5	3.33 (.58)	.67 (1.15)
9	7	4.67 (1.15)	.00 (.00)

10	1	.00 (.00)	.67 (1.15)
----	---	-----------	------------

As we noted earlier, our attempt to clarify data collection for SA level 2 increased the number of risk indicators that subjects had to check. Referring to Fig. 4, a subject identifying a car ahead and to the right would check (under “There are Vehicles”) “ahead” and “in the right lane.” Subjects could explicitly check the “There are Vehicles” box but were automatically given credit for it if they checked any of the boxes in that category.

As in experiment one, subjects had more omissions than additions. There were clearly more problems with SA level 2 than in our previous experiment. Not surprisingly, the scenarios that had more risk indicators had the most omissions.

For the SA level 3 we asked for a rationale for their choice. Table 8 shows the responses of subjects to SA level 3.

Table 8. The results for SA level 3 from experiment two

Scenario	% Correct	% Didn't know	Correct Rationale (# of responses)
1	33		0 (2)
2	100		3 (3)
3	100		3 (3)
4	33	67	1 (2)
5	100		1 (3)
6	100		2 (2)
7	67		2 (3)
8	67	33	0 (2)
9	67	33	1 (2)
10	67		1 (2)

Scenario 4 was the most problematic. In this scenario, we asked if the car could safely move into the left lane. The correct answer was no as there was a car in that lane. One subject answered correctly and provided the correct rationale. The other two subjects did not know. Scenario 4 also had a high number of omissions for SA level 2. Therefore, it was not surprising that subjects had problems with SA level 3. In scenario 5, all subjects answered correctly, but only 1 had the correct rationale.

XII. EXPERIMENT THREE

In our third experiment, we tried a different approach. We kept the data collection method for the SA levels the same as in experiment two. We changed the design of the user interface to a more graphical design of the supervisory interface. Although we were not pleased with the elicitation method for SA level 2 information, we needed to make sure that we varied only one condition at a time. This interface design was motivated by comments from our subjects in the two previous experiments. They commented that viewing the situations graphically seemed more intuitive to them. In this presentation method, the scenarios were animated and presented in the lower right hand corner of the display. The icons were still shown on the view of the map but the more detailed text descriptions were now replaced by the graphical

animation. The speed and fuel information was also displayed in a more graphical fashion. The redesigned user interface is shown in Fig. 5.

A. Subjects

Eight subjects participated in this experiment. All eight held a valid driver’s license. Six of these subjects were male and two were female. None of these subjects had participated in experiment one or experiment two.

B. Experimental Design

We gave the subjects three training tasks to familiarize them with the procedure and then administered the ten tasks, using the same SA level 1, level 2 and 3 questions as in experiment two. We used the refined collection user interface shown in Fig. 4. We did not administer the post-experiment questionnaire for these three subjects.

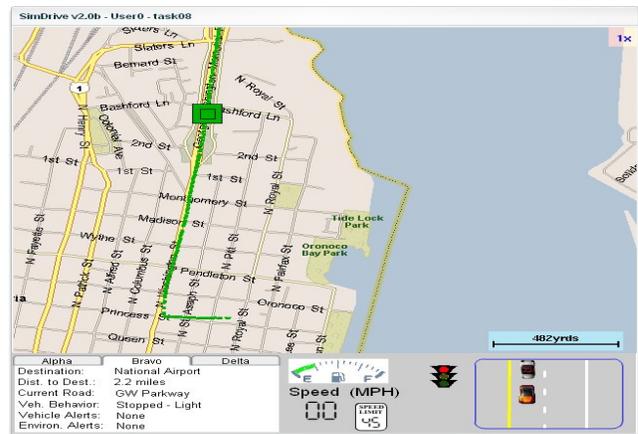


Figure 5. Graphical user interface used for experiment three

C. Results

Table 9 shows the results for SA level 1 from experiment three. A blank cell in Table 9 indicates that all responses were correct.

Table 9. The percent of incorrect responses for the 3 categories of SA level 1 for experiment three

Scenario	Vehicle	Environment	Route
1			12.5
2	12.5		
3			25
4			37.5
5		12.5	25
6			12.5
7	25		
8		25	25
9		12.5	
10			

As in experiments one and two, the route information was more problematic than the vehicle and environment status.

The display of the route information had not changed so we did not anticipate improvements in these responses.

Table 10 shows the responses for SA level 2 given the more graphical presentation of the information. We made 2 changes to the number of risks from experiment three. Scenario 4 had the debris moved to the front of the vehicle which decreased the number of boxes that subjects needed to check by 1. A car and debris were added to scenario 8 increasing the number of risk boxes to be checked by 4.

Not surprisingly, there are more omissions for scenarios with a higher number of risks. There are also more additions in scenarios with larger numbers of risks. Scenario 6 seems to be an exemption to this. Scenario 6 was extremely boring in that no risks changed during the scenario.

Table 10. SA level 2 responses for experiment three

Scenario	# of risk indicators	omissions– Mean(SD)	additions– Mean(SD)
1	2	.38 (.52)	.00 (.00)
2	1	.00 (.00)	.00 (.00)
3	6	1.13 (.64)	.50 (.53)
4	6	1.00 (1.31)	.25 (.46)
5	3	.38 (.52)	.25 (.46)
6	5	.50 (.93)	.00 (.00)
7	2	.88 (.83)	.13 (.35)
8	9	2.75 (1.16)	.25 (.46)
9	7	2.13 (1.36)	.13 (.35)
10	1	.00 (.00)	.00 (.00)

Table 11 shows the responses for SA level 3.

Table 11. The results for SA level 3 from experiment three

Scenario	% Correct	% Didn't know	Correct Rationale (# of responses)
1	75	25	4 (5)
2	100		7 (7)
3	100		8 (8)
4	87		8 (8)
5	100		8 (8)
6	100		7 (8)
7	100		7 (7)
8	87		7 (8)
9	100		8 (8)
10	87		7 (8)

Table 12. Post-experiment questionnaire for experiment three

Question	Mean rating (1 – extremely easy, 7 – extremely difficult) (SD)
Determining overall condition of vehicle	1.6 (9.5)
Determining overall condition of environment	1.75 (0.5)
Determining overall condition of route	2.25 (1.0)

Awareness of situation (1 – always aware, 7 never aware)	2.25 (0.7)
How many cars do you think you could monitor with this UI?	2.6 (ranged from 2 to 4) (1.1)

Although we asked for a rationale, not all subjects gave one. Thus column 4 shows the correct rationale and the number of responses we received for that scenario. Interestingly, for scenario 4, one subject gave the correct rationale but answered the questionnaire incorrectly. This could have been just a slip.

Subjects felt it was easy to determine the condition of the vehicle and the environment, more so than the condition of the route. They felt they could monitor several vehicles using this type of user interface. Table 12 contains the results of the post experiment questionnaire we gave subjects.

XIII. DISCUSSION

Experiment three shows an improvement in SA level 3 understanding over that measured in experiments one and two. There were fewer “I don’t know” answers and a greater percentage correctly answering the question. We also saw an improvement in SA level 3 between experiments two and three. However, as we had so few participants in experiment two, we cannot draw any conclusions from this.

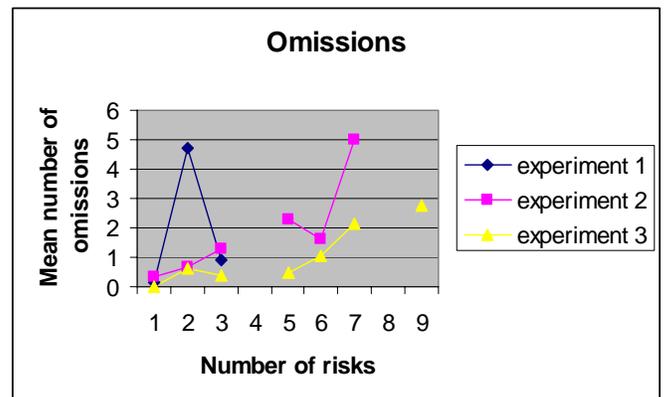


Figure 6. Comparison of number of risk indicators and mean omissions by experiment

Fig. 6 shows the mean omissions for the number of risk indicators for the three experiments. That is, for a given number of risks (1-9), how many omissions did subjects average in the different experiments? Recall that the change we made to elicit SA level 2 responses more precisely increased the number of risk indicators that subjects needed to mark. Therefore, experiment one had no more than 3 risk indicators, while experiment two had a maximum of 7, and experiment three had a maximum of 9, although we had no scenarios with 4 or 8 risk indicators. For 1 to 3 risk indicators, subjects in experiments two and three performed about the same and both performed either better than or similar to subjects in experiment one. For higher number of risk indicators, subjects in experiment three performed either better or approximately the same as subjects in experiment two. However, due to the small number of subjects in experiment two, this data must be treated with caution.

However, when combined with the results for SA level 3, we are able to say that SA level 2 is improving as SA level 3 cannot be achieved without first gaining SA level 2.

We believe that the way we elicit SA level 2 needs to be refined considerably given the number of omissions by subjects. We need to do several things. We need to analyze our assessment questions to determine if the factors we ask about for SA level 2 are necessary and sufficient to answer the SA level 3 questions. That is, we need to determine not only overall results, but which risk indicators in particular are being missed. Secondly, we need to look at the order in which subjects fill out the SA assessment questions. If the majority of the subjects fill them out in order, then we can try varying the order of questions to determine if the failure to answer correctly is a limitation of short term memory. We also need to look at the particular items that subjects miss and analyze the scenarios to determine how long the risk has been visible to the subject prior to the assessment. This would help determine the length of time that is needed for subjects to see and comprehend SA level 2 information. Being able to vary the lengths of time subjects have to recognize and comprehend risks and measuring the resulting SA level 2 would be a valuable evaluation technique for supervisory human robot interfaces.

We are considering experimenting with a more visual approach to collecting data to assess level 2 SA. We could present a set of graphic representations similar to those presented in the human-robot interface to determine the situation awareness. This presents an interesting question about the relationship between the presentation of the information in the human robot interface and the data collection tool.

XIV. CONCLUSIONS AND FUTURE WORK

We believe that the method we developed based on the SAGAT methodology for assessing situational awareness shows promise for assessing supervisory user interfaces. The three experiments we conducted lead us to believe that the methodology is repeatable and does have the sensitivity to discriminate between user interface treatments.

However, developing the assessment questions, even for a domain where expert information is readily available, is difficult. We think that the data collection tool we used is a reasonable starting point for assessing SA level 1 and SA level 3. We need to find a better way to assess SA level 2. Participants in our experiment had difficulty as the number of risks increased. This is problematic as we have not yet considered the more complex scenario of driving on urban streets with many pedestrians, parked cars, and children playing. We need to examine the tension between precision in data collection and increasing the amount of information the subject needs to recall in order to perform well. We also need to determine how the amount of time users have to monitor the different conditions affects performance.

We also need to determine a way to validate our methodology. In a separate field study of off-road semi-autonomous driving vehicles, we calculated the time vehicle

operators needed to acquire situation awareness when an intervention was requested by the vehicle [11]. In this analysis, we measured the time that operators spent manipulating the remote vehicles' cameras, prior to either taking control and teleoperating or issuing a command to the vehicle. Not surprisingly, the more specific the intervention request, the less time it took operators to acquire SA. We also found differences due to different types of terrain [12,13]. We are considering devising an experiment to investigate the difference between user interfaces using SA acquisition time as the metric. It would be interesting to run the same scenarios as in the experiments reported in this paper, but asking subjects to intervene at the same point as when we froze the simulation. Subjects would be asked to press an "action" button when they were ready to "take control." We would record the time this took in addition to asking subjects to select a plan from a multiple choice set. We could then determine if there was a relationship between SA acquisition time and the results from our SAGAT study.

Our long term intent is to develop supervisory user interfaces for multiple vehicles. The SAGAT methodology will have to be modified to accommodate multiple vehicles. For this domain, we will need to seek out experts in similar areas, such as air traffic control and monitoring of public transportation. We will need to develop the appropriate level of abstraction, both for the SAGAT method and for the human-robot interface.

If a number of robots and humans are working as a team, we will need to develop methods to assess the situation awareness of all members of the team [14, 15,16,17]. Team members, both humans and robots, will need to be aware to some extent of what the others are doing. We will also need to extend this to accommodate the notion of roles [3]. A supervisor of multiple robots will need to be aware of when operators are assisting some of them or when robots are part of a team and may be given instructions by a peer.

Our research at this point is still exploratory. We need to conduct more experiments, first to refine our implementation and eventually to test the reliability of our work. We believe that development of such a methodology is necessary for assessing how well human-robot interfaces facilitate the acquisition of situation awareness.

For those interested in applying this SA methodology, the simulation, the user interface designs, and the domain questions are all available by contacting the authors.

ACKNOWLEDGMENT

The authors want to thank Paul Hsiao for his work in data collection. Also, we want to thank the many colleagues who participated in our experiments. We thank the reviewers for their many helpful comments.

REFERENCES

- [1] ISO 9241-11, (1998) "Ergonomic requirements for visual display terminals, usability guidance" [Online] available at International Organization for Standardization, <http://www.iso.org/iso/en/ISOOnline.frontpage>.

- [2] J. Scholtz, "Theory and Evaluation of Human Robot Interactions", in *Proceedings of Hawaii International Conference on System Science.(HICSS 36)*, 2003.
- [3] J. Scholtz, "Creating Synergistic CyberForces" in *Multi-Robot Systems: From Swarms to Intelligent Automata*, A. C. Schultz and L. E. Parker, Eds, Kluwer, 2000.
- [4] M. Endsley, "Theoretical Underpinning of Situation Awareness: Critical Review" in Mica R. Endsley and Daniel J. Garland (Eds.) *Situation Awareness Analysis and Measurement*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 2000. 3-32.
- [5] A. Pritchett and R. Hansman, R. "Use of Testable Responses for Performance-Based Measurement of Situation Awareness" in *Situation Awareness Analysis and Measurement*, M. R. Endsley and D. J. Garland, Eds. Mahway, New Jersey: Lawrence Erlbaum Associates, 2000, pp.189-209.
- [6] K.A. Ericsson and H.A. Simon, *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA. MIT Press. 1993.
- [7] D. Jones, "Subjective Measures of Situation Awareness" in *Situation Awareness Analysis and Measurement*, M. R. Endsley and D. J. Garland, Eds. Mahway, New Jersey: Lawrence Erlbaum Associates, 2000, pp. 113-128.
- [8] M. Endsley, "Design and Evaluation for situation awareness enhancement". In *Proceedings of the Human Factors Society 32nd Annual Meeting*, 1988. vol. 1, 97-101
- [9] GPS-Drive [Online] available at <http://gpsdrive.kraftvoll.at/index.shtml>.
- [10] GNU General Public License, [Online] available at <http://www.gnu.org/copyleft/gpl.html>
- [11] J. Scholtz, B. Antonishek, B., and J. Young, "Evaluation of Operator Interventions in Autonomous Off-Road Driving". In *Performance Metrics for Intelligent Systems, PERMIS 2003*. 2003.
- [12] J. Scholtz, B. Antonishek, and J. Young, "Operator Interventions in autonomous Off-road Driving: Effects of Terrain" in *Proceedings of System, Man, and Cybernetics*, 2004.
- [13] J. Scholtz, 2004. "The Effect of Situation Awareness Acquisition in Determining the Ratio of Operators to Semi-Autonomous Driving Vehicles", in *Proceedings of SPIE Interaction Symposium: Optics East*, 2004.
- [14] J. Drury, J. Scholtz, and H. Yanco, "Awareness in Human-Robot Interactions". In *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics*.
- [15] C. Bolstad and M. Endsley, "Shared Mental Models and Shared Displays" in *Proceedings of the 43rd Meeting of the Human Factors and Ergonomic Society.*, 1999.
- [16] N.J. Cooke, E. Salas, J.A. Cannon-Bowers, and R. Stout, "Measuring Team Knowledge", *Human Factors*, 42, 2000, pp. 151-173.
- [17] C. Bolstad and M. Endsley, "The Effect of Task Load and Shared Displays on Team Situation Awareness" in *Proceedings of the 14th Triennial Congress of the International Ergonomics Association and the 44th Annual Meeting of the Human Factors and Ergonomic Society*, 2000.



Jean C. Scholtz (M'92) received the B.A. degree in mathematics from the University of Iowa, Iowa City, IA in 1966, the M. S. degree in mathematics from Stevens Institute of Technology, Hoboken, NJ, 1969, and the Ph.D. degree in computer science from the University of Nebraska, Lincoln, NE in 1989.

She is currently a Computer Scientist at the National Institute of Standards and Technology, Gaithersburg, MD. She was previously a Program Manager in the Information Technology Office at the Defense Advanced Research Projects Agency (DARPA). She has held positions at Intel Corporation and was on the computer science faculty at Portland State University, Portland, OR. She also worked at Bell Laboratories, Murray Hill, NJ. She is the author of over 50 papers in books, professional journals, and conference proceedings. Her research interests include human-robot interaction and evaluation of interactive systems. She has held positions on the SIGCHI executive board and is active in the CHI conferences. She is on the editorial board of *Empirical Software Engineering*, *Interacting with Computers*, the *International Journal of Human Computer Studies* and *ACM Interactions Magazine*.

Dr. Scholtz was awarded a NASA/ASEE summer fellowship in 1991 and 1992 for work at the Kennedy Space Center. She is a member of ACM, SIGCHI, AAAI and UPA.



Brian Antonishek received the B.S. degree in computer science from the University of Pittsburgh at Johnstown, Johnstown, PA in 1990 and the M. S. degree in computer science from the University of Pittsburgh, Pittsburgh, PA in 1996.

He is currently a Computer Scientist at the National Institute of Standards and Technology, Gaithersburg, MD. He was previously a Software Engineer at Concurrent Technologies Corporation, Johnstown, PA. His research interests include robotic control software and human-robot interaction.



Jeff D. Young received a B.S. degree in Industrial Psychology from Abilene Christian University, Abilene, TX in 2002, and a M.A. degree in Human Factors and Applied Cognitions from George Mason University, Fairfax, VA in 2004.

He is now at Resource Consultants, Inc. Vienna, VA. Previously he was an Operations Research Analyst at the National Institute of Standards and Technology, Gaithersburg, MD. His research interests include human-robot interaction and usability in critical systems. He is a member of HFES, SIGCHI, and ACM.