A methodology for the assessment of 360° Local Area Awareness displays

Chris Manteuffel[a], Jason Metcalfe[a], Tony Johnson[a], Matthew Jaswa[a], Xung Nham[a], Brad Brumm[b]
[a]DCS Corporation, 6909 Metro Park Drive, Suite 500, Alexandria, VA, 22310
[b]U.S. Army Tank, Automotive, Research Development and Engineering Center, 6501 E. 11 Mile Road, Warren, MI 48397-5000

ABSTRACT

In the process of developing new technologies for displaying 360° visual data supporting Local Area Awareness (LAA) in complex environments (e.g. tactical military environments), one important, though often overlooked, area is system evaluation. Without an accurate and reliable evaluation, it is impossible to determine which elements of the new display are useful and which need further development. Evaluating a system properly requires two types of tests: one for testing capabilities (e.g. given a display, what types of threats can be detected and identified?), and another for probing whether a given display configuration is useful (e.g. will the human operator use this more complex interface appropriately in the real world?). While established methodologies exist for the former, the latter often appears as a much less tractable problem. This is primarily because of the difficulties with modeling the complexity of the real world in a simulated environment. This paper presents a methodology for architecting a distributed simulation to support evaluation of a 360° LAA display system for usefulness to human participants within virtual environments. The evaluation that leveraged the methodology ultimately reported several unexpected results due to the effectiveness of the evaluation; for example, the experiment discovered a much greater "keyhole effect" than expected, where participants focused almost entirely on the forward 180°, even when presented with imagery covering the full 360°. Such results demonstrate the utility of the methodology, particularly for developing evaluations that discover unexpected aspects of operational use in complex environments.

Keywords: 360 visualization, experiment design methodology, local area awareness, system evaluation

1. INTRODUCTION

The U.S. Army has been undergoing rapid transformation and modernization since the beginning of the current millennium. Formerly undertaken as part of the Future Combat Systems program, advanced technology efforts, including those in robotics, network science, sensors and others, are c under the current Army Brigade Combat Team Modernization Program[1]. Among the areas of focus that are currently of high priority within these Army science and technology efforts is that of evolving Intelligence, Surveillance and Reconnaissance (ISR) capabilities[2]. Taken with the continued development of battlespace network resources, sensor technologies aimed at providing full-spectrum, real time Local Area Awareness (LAA)[3], and implementation of adaptive automated systems, such technologies hold great promise for providing Soldiers with the information necessary to achieve and maintain greater situational awareness and reduced workload while performing mission-critical tasks such as local area awareness while remaining secure inside the vehicle[4]. Such enhanced information acquisition and utilization capabilities should lead to improved lethality and survivability of U.S. Troops.

Yet, with an accelerated pace of technology development and integration comes a greater burden of responsibility for those tasked with evaluating and demonstrating the safety and effectiveness of new systems. That is, the assessment of system function and usability represents an essential step in the process of transitioning science-based technologies into the hands of the Soldiers in the field[e.g. 5]. Moreover, the task of designing and executing evaluations to keep pace with the rate of technologic expansion in the modern era, particularly when the availability of testbed systems and field test venues is limited, represents an important challenge to systems, software, and human factors engineers who work diligently to validate newly advanced technologies. These challenges make effective testing all the more important

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
25 APR 2011	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A methodology for the assessment of 360° Local Area Awareness displays	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)	5d. PROJECT NUMBER
Chris Manteuffel; Jason Metcalfe; Tony Johnson; Matthew Jaswa; Xung Nham; Brad Brumm	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
US Army RDECOM-TARDEC 6501 E 11 Mile Rd Warren, MI 48397-5000, USA DCS Corporation, 6909 Metro Park Drive, Suite 500, Alexandria, VA 22310	**21664RC**

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
US Army RDECOM-TARDEC 6501 E 11 Mile Rd Warren, MI 48397-5000, USA	**TACOM/TARDEC/RDECOM**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) **21664RC**

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at SPIE 25-29 April 2011 Orlando, Florida, USA, The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**SAR**	**14**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

because they also impede determining which elements of the new systems are useful and which need further development.

Minimally, the proper evaluation of a new technology or capability requires two general types of tests: those assessing utility and those assessing usability[6,7,8,9]. The first type of evaluation can be considered as that of testing capabilities and basic functional system characteristics. An example of this type of test would be to determine what types of objects and entities could be detected and identified with a particular display system at a given range and using a particular field of view. A second type of evaluation can be thought of as probing whether a given system is useful to the end users of the system. This second type of test, however, represents the challenge of examining how the system would be used by trained operators *within the appropriate context*[10]. An example of such an evaluation could involve testing whether a given user interface for an advanced display system is useful or whether the human operator would deploy the more complex interface capabilities successfully in the real world.

While established, and often domain-specific, methodologies exist for functional types of testing (e.g. examinations of sensor resolution, network throughput, etc.) assessment of system usability in context often appears as a much less tractable problem. In large part, this difficulty arises from the uncertainty involved in modeling the complexity of the real world with a fidelity that affords evaluators the resolution necessary to make precise, reliable, valid and generalizable interpretations of human behavior; especially challenging is when such inferences need to hold across variable operational contexts (military examples include the type of diversity currently seen in U.S. overseas operations, such as urban versus rural environments). As a means of framing discussion of these types of problems, the present paper outlines a systems approach to assessing a 360° LAA display system that was designed to enhance the situational awareness of Soldiers conducting operations from within completely armored vehicles (such as the popular and heavily used M1126 Stryker Infantry Carrier Vehicle). The specific environment in which the 360° display system was intended to be applied was that of a military threat zone in an urban setting. The discussion focuses in large part on the techniques used to provide a simulation of the military operational context that allowed for valid interpretation of user preferences and performance with a variety of possible display configurations. Data are presented as exemplar outcome variables from the modeling and simulation efforts.

2. METHOD

2.1 Apparatus and Software

The assessment used a simulation engine to test a Warfighter-Machine Interface (WMI) developed for the Improved Mobility and Operational Performance through Autonomous Technologies (IMOPAT) Army Technical Objective (ATO). The interface was designed to allow a single Soldier to monitor the full 360° around the vehicle. There were four different display configurations under examination, shown in Figure 1. The first presented the feed of a single fixed camera (one of six located around the vehicle) at a resolution of 768x1024 pixels (Display A).The sensor portal, as it was called, was common to all display configurations and was controlled by the user selecting which fixed camera feed to view. Another display presented the three forward cameras in a banner display across the top (reduced in such a way as to keep the aspect ratio unchanged- for a total of 1728x369 pixels) and then had room for the same size of display as in the first option (Display B). Unlike the portal, the banner in B and C was not controllable and provided a fixed view of the forward 180° at all times. In Display C the banner was reduced to 1248x369 pixels and in this case there was some loss of vertical field of view, while the sensor portal remained unchanged. This was the same type as B, but reduced to allow for small widgets along the left side of the screen to display information a vehicle commander might need to conduct a militarily relevant mission (to keep the experiment controlled, no information was presented in this area). The final view (Display D) had two banners, with the top one showing the forward 180° and the bottom one the rear 180°, allowing the Soldier to view all six cameras at one time, and had a central portal giving them a view from a single camera at a reduced resolution of 768x576. The four displays represented different points along the spectrum of 360° LAA display screens: in general one has a choice between presenting a wider field of view with less image resolution- and hence a loss of sighting at a distance- or giving better range performance at a loss of field of view.

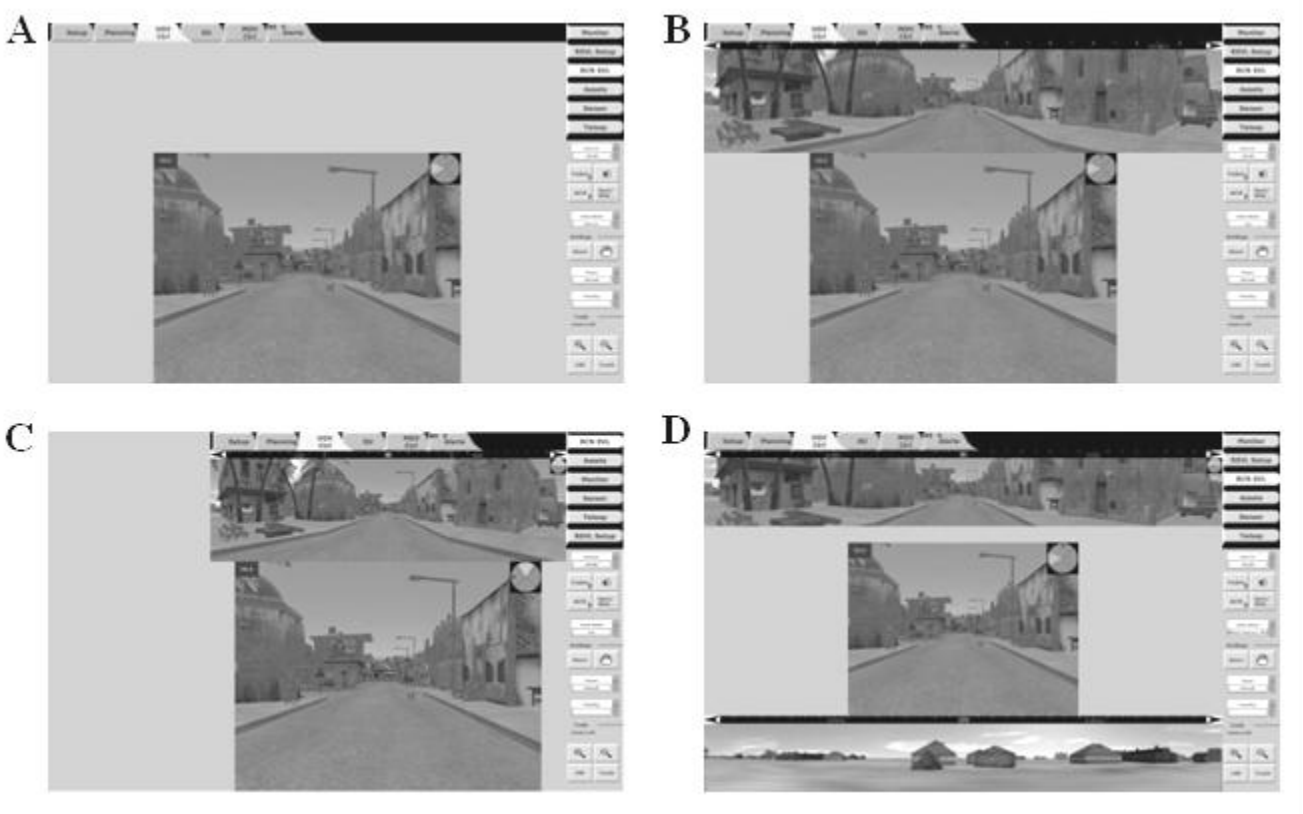

Figure 1. The four LAA Displays Under Evaluation

One aspect of the experiment to take advantage of simulation was that each banner was a cleanly "stitched" image with no seams from each virtual camera. This was feasible in the simulation because each camera was a single point, so it was easy to have all three cameras on top of each other for a single, perfectly aligned image. Using actual sensor technology, where cameras occupy volume, this is not possible without a lot of very computationally intensive software running in hard real time.

In addition to the displays, a data capture widget was added to the WMI. The widget was invoked as a button that the participants were to push when they identified a threat. Once enabled, the widget presented the participant with four buttons: three different threat types and a cancel button. When one of the first three buttons was selected the Soldier was presented with 12 buttons, allowing them to report the direction of the threat, relative to the vehicle heading in terms of clock-position (a standard military technique, with 12 o'clock as straight ahead of the vehicle and 3 o'clock as directly to the right of the vehicle).

Shown in Figure 2 below, a total of four computers, running two different operating systems and running 8 different components created by the experiment team, were necessary to run the distributed simulation environment, WMI and data acquisition systems used for this experiment

The simulation engine would simulate a vehicle with suitable cameras for the system, drive the vehicle, and present potential targets and distractions to the participants. It was composed of several components split across two computers. One component, the Embedded Simulation System (ESS), actually modeled the participant's vehicle, and did the necessary work to drive the vehicle along the a priori route. Another component, the Event Server, tracked the vehicle as it followed the route, and activated other components at the correct time. The three components the Event Server would activate were the Scenario Populator that was responsible for placing all the targets and entities within the environment, a data logger to record when events were activated, and a sound player to play pre-recorded sound files announcing high value targets, etc. at the appropriate moments.

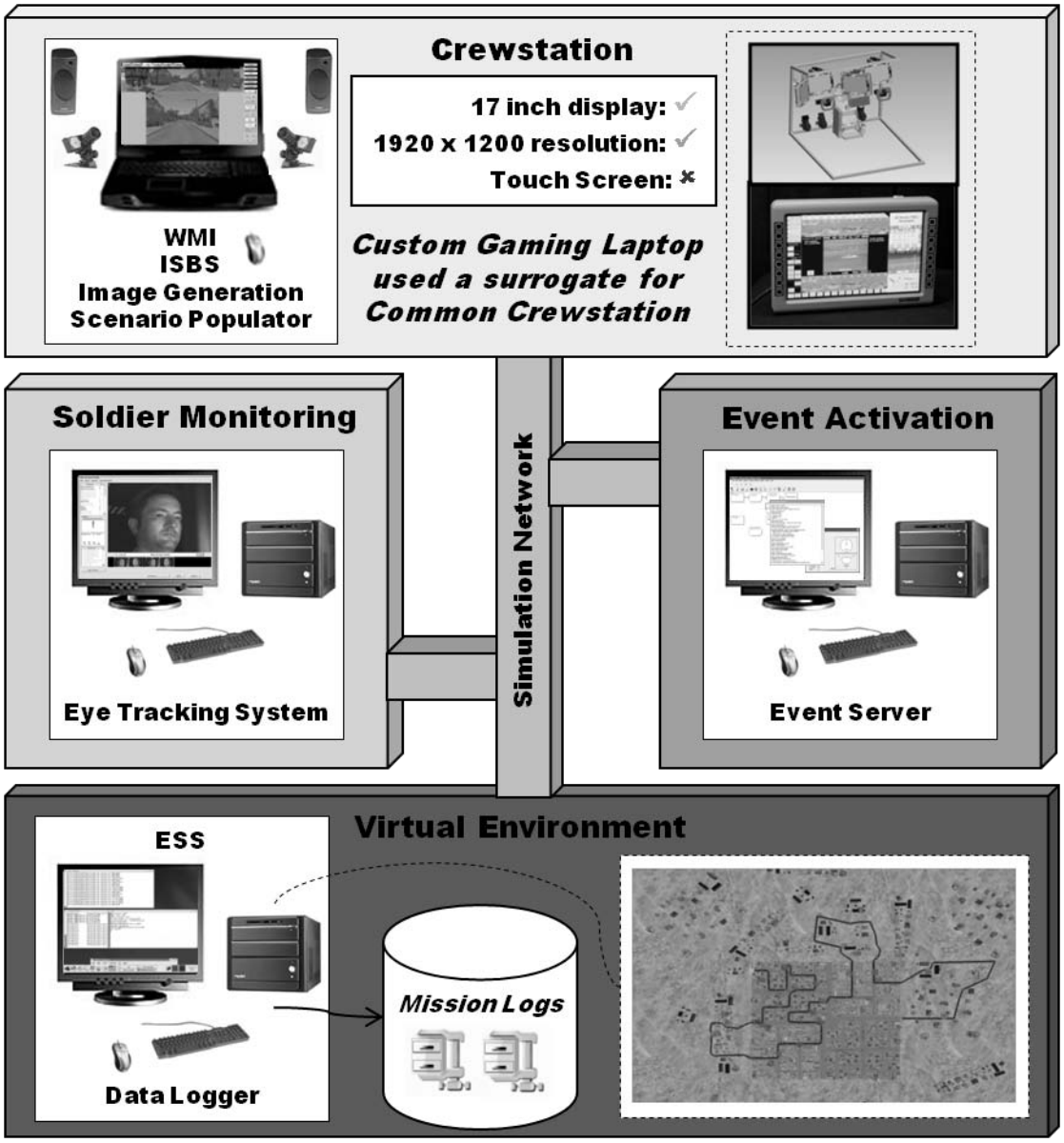

Figure 1. Software Configuration

A final component of the simulation engine was the DIS Recorder. The Scenario Populator would broadcast Distributed Interactive Simulation (DIS) protocol packets to the WMI's Image Generator, and this component would record every packet sent over this protocol. It thus provided the ground truth for every run: the exact location of every entity at every moment, so that the experimental team could trust that it had the correct data logged.

One component was not created by the team: commercial eye-tracking software. It used infrared (IR) video cameras to track the reflections of the eyeballs to track eye gaze direction. It recorded data at 60 Hz to a separate machine. There were two problems with this setup. One was that glasses simply defeated the system. Between IR reflections, frames blocking the camera, glare, and simply more complex features, the system was unable to effectively track the gaze of anyone wearing glasses. A second problem was that, because the team only had access to two cameras, there was minimal ability to handle head movement. The team was unable to obtain a third or fourth camera, which would have given a wider field of view suitable for dealing with head movement. Instead, whenever the head moved from where the

cameras had been calibrated for, the eye tracking software lost track of the gaze. This limited the effectiveness of the eye-tracking, though it still provided useful explanation of unexpected results.

There were two more tools developed by the experiment team. One was the Data Analysis and Reduction Tool (DART). This was the automated tool discussed below in the Data Analysis section. A final tool developed by the experiment team allowed determining the location of targets, distractions, and routes in the environment and generated all the scenario specific configuration files necessary to run all the various components of the distributed simulation engine.

2.2 Experiment Design

Figure 3 below represents the overall systems and software engineering approach that was used to evaluate the 360° LAA system. A critical element of the methodology, which dominates discussions with the experimenters at the outset of planning an evaluation, is a hierarchy of questions that need to be answered. First, what is the basic research question? Then, what metrics are to be used? Finally, what do the scenarios look like? After those answers are refined, the scenarios can be properly implemented. After a verification of the entire system – including the data capture process – the experiment is ready to collect data from actual participants. Then a data processing and analysis phase commences, and then the results of that process are fed back into the requirements for the follow-on experiments. For this experiment the most important question was whether the displays improved performance on local security tasks over that observed with current force technologies. Given that basic research question, the team then had to determine what the best metrics were for answering it, and then they had to decide what sort of scenarios would support those metrics.

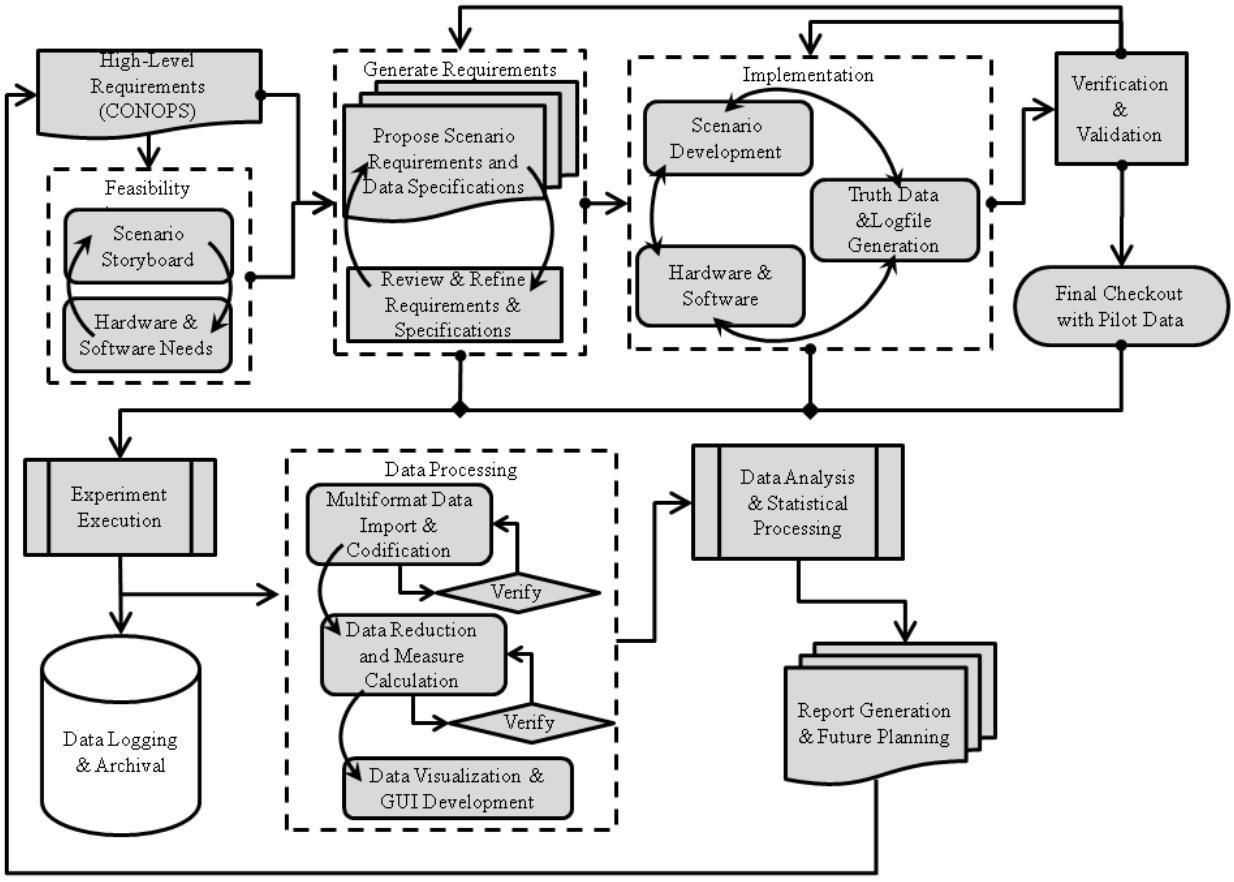

Figure 3. An overview of the methodology of the experiment development process. Figure from Jason S. Metcalfe, Gabriella Brick Larkin, Tony Johnson, Kelvin Oie, Victor Paul, and James Davis, "Experimentation and evaluation of threat detection and local area awareness using advanced computational technologies in a simulated military environment." Unmanned Systems Technology XII, Grant R. Gerhart, Douglas W. Gage, Charles M. Shoemaker, Editors, Proc SPIE 7692, 769209 (2010).

In discussion with the program scientists, the experiment team identified a set of metrics that were necessary to describe the behavior of the participants and to differentiate between levels of effectiveness. The easiest to capture was subjective appraisal of workload, which was captured after each run by the use of a NASA-Task Load Index (a multi-dimensional rating survey)[11]. Another metric that was easy to capture was the set of user preferences as to interface design, which was collected by means of an exit interview conducted after the full set of experiment runs.

Neither of the above qualitative (and subjective) methods were considered particularly reliable; in the experience of the experimental team, users are often surprisingly wrong about their own experiences. Therefore, quantitative metrics were examined as well. One such metric was the proportion of threats detected. In general, it was expected that a better 360° LAA interface would allow a Soldier to spot more targets in the environment. Tied to that metric was the accuracy of threat identification. Participants were required to report both whether the target was armed or unarmed and its location relative to the vehicle heading. These two metrics provided some objective ability to validate participants subjective self-reports.

The most important metric, in terms of influencing the structure of our evaluation, was reaction time. While typically used in more narrowly focused experiments in psychology, reaction time was measured in the present experiment as the time elapsed from the moment when a target came in view of any camera (regardless of whether the operator was looking at that camera view) until the moment when the participant started a simplified digital report. This gave the research team a useful measure that was objective and could be compared easily, allowing for gradations in capability to be observed. It also added a temporal dimension to the interpretation of threat identification performance. That is, as the results of the study bore out, evaluation of reaction time appeared to reveal cognitive-behavioral strategies that the participants were using as they responded to the targets presented in the environment.

Based on the identified experimental requirements, the team concluded that 38 target presentations would be necessary on each trial. This would enable the recording of an adequate number of observations to give the statistical power necessary with the number of participants that were available. For running time purposes, it was necessary to have each trial conducted inside of 20 minutes – any longer and there would be fatigue issues with participants, and the team was concerned about scheduling participants if they were needed for more than half a day.

2.3 Scenario Development

Once the basic question of the metrics was decided upon, the scenario development process could begin. Scenario development was by far the most time-consuming part of preparing the experiment, but it was absolutely vital for the success of the experiment that it be done in an orderly, considered manner. The methodology for scenario development started with selecting events that would take place, then building an appropriate database to support those events, then creating the vehicle route, then matching events to locations and laying them out. With a new observation required approximately every 30 seconds (20 min trials / 38 target presentations), and because of the dynamics of the experimental scenario (involving both a simulated moving vehicle as well as a large variety of both static and dynamic human and non-human entities), the experiment required a great deal of precise timing and coordination. A significant number of iterations were necessary to get those correct. That is, in order to ensure that the experimental participants had a fair chance to observe each and every target, the scenario developer would place the entities within the database and then run through the scenario to ensure that events were not occluded by the environmental dynamics. For example, it was possible that a target could be moving in a manner that it was occluded by some environmental feature, such as a wall, a large stationary object or even another similar entity (one human blocking another), due to the perspective created by its motion and that of the vehicle. In such a case, the scenario developer would adjust the position of the entity – perhaps changing its starting position or velocity– and then run the scenario again to ensure that the occlusion was removed. In the current experiment, four unique, though statistically similar scenarios were used and there were 38 target presentations per scenario (excluding distracter entities, of where there were approximately 60 – 70 per scenario). Therefore, many iterations were necessary to get the placement and timing to be both repeatable and correct across all 152 target presentations.

In order to increase the ecological validity of the system assessment, the research team initially spent a great deal of time trying to determine different environmental elements (i.e. threatening human behavior as well as objects that could serve as weapons such as IEDs) that would seem naturally threatening to Soldiers. To do this, the team performed significant background research involving multiple sources of information including: interviews with subject matter experts like Soldiers and other researchers who had gone on trips to the battlefront, web logs authored by Soldiers who had been (or

were currently) in-theater, press releases regarding military activities, Army field manuals, and contextually-appropriate pictures and videos available on the Internet.

After much discussion focused on how to present potentially threatening events in the simulation engine, a set of possible hostile behaviors were chosen. A salient example of a behavior selected to identify an entity as a threat was that of brandishing a weapon; any entity doing so was automatically considered a potential threat and would need to be reported. Of course, using such an obvious visual cue was considered just one of many ways that humans would be apparent threats to Soldiers in context. Indeed, 360° LAA display screens, if effective, should enable Soldiers to detect more subtle differences between threatening and non-threatening individuals. In order to evaluate what screen designs were superior from this perspective, the team had to examine how well the participants could understand all of the cues presented within the environment surrounding their vehicle, as well as to understand the significance of entity behaviors and draw conclusions from these behaviors over time. From a scenario design standpoint, this meant that there had to be behaviors and patterns that marked unarmed people as potential threats.

One example of a particularly subtle threatening behavior that was selected was someone standing still and staring directly at the vehicle as it passed – such a behavior was chosen to represent a "spotter" who would typically be waiting to signal detonation of an Improvised Explosive Device (IED). The team also took advantage of radio messages to force threat reports. That is, the participants would hear a report warning them to watch for a certain thing, such as a car, a broken TV (fashioned into an explosive), or a person dressed in a certain way. Sometimes, the participants would later see something matching the radio description and thus be required to report on the observed entity. In order to remain a test of interface displays rather than memory, if a participant was going to see something that was to be reported in this manner, it was presented within 30 seconds of the radio alert. Of course, only the experimenters and scenario designer were aware of this time window; the experimental participants did not know that the entity, were it to appear, would do so within a limited period of time- and ¼ of all warnings were not associated with any entity in the environment. To test the effectiveness of the displays at spotting IEDs, bits of rubble or trash were strewn about the landscape and wires (representing command wires) were attached to some of them. Items with such wires sticking out were considered to be IEDs and it was expected that the participant would report on them.

With the threat set settled, the team turned to the issue of creating an appropriate terrain database for the scenarios. Because one of the design requirements for the 360° LAA display system was to provide man-sized target detection capability out to 220 meters from the vehicle, virtual cities were not sufficient to test the full capabilities. Urban areas simply do not provide enough vistas out to 220 meters to provide a meaningful analysis at range. To account for this issue, a suburb area surrounding the urban core was constructed as a series of built up areas with large intervening spaces. These outer buildings gave the participants an opportunity for quite long sight lines, thus allowing an assessment of the system at the intended range. Because the 360° LAA displays were designed to support threat detection but not classification (lacking the ability to zoom in and closely interrogate specific elements of the environment, for example) it was felt that requiring participants to differentiate between armed and unarmed targets at range would be too difficult. That is, it would be impossible to tell if they could detect a person but not his weapon at range. So, while in the urban core of the database participants were expected to report only on suspicious people and ignore non-threatening people, participants were instructed to report on everyone during their time in the suburbs. The justification for this was that the suburbs were under curfew. In this, we have an example of how scenario design and the narrative presented to participants must conform with one another. This need is particularly an issue when some of the experimental participants come from a specialized population, such as Soldiers, that have specific expectations regarding contextual explanations for observed events. While breaking this context is perfectly acceptable for tests of elemental system performance, it degrades the usefulness of the system usability tests.

As stated earlier, the reaction time metric required special attention and exerted particular influence over the scenario design. Specifically, in order for the time to be meaningful, it was necessary to avoid presenting two targets at once; doing so would unfairly penalize the reaction time for the second target while the participant filled out a threat report for the first one. This meant that the route needed to wind through much of the city as well. That was because the route needed to involve a lot of turns inside a city as each turn removed an old target sighting and consequently brought a new target sighting opportunity. With the requirement that a new sighting appear once every 30 seconds (approximately) the density of turns and short sight lines of a city constituted a design requirement. In the end, a 2/3 in the city, 1/3 in the suburb split was chosen as the best compromise between allowing reporting opportunities on long range targets and yet providing enough total number of targets necessary to draw meaningful quantitative conclusions.

Four 20 minute routes, each with 12 minutes in the urban areas and 8 in the suburban areas, were created. In order to ensure the precise timing necessary for such quick target presentations and to allow for the clearing out of old targets, the vehicle the participants would (virtually) ride in was controlled by an autonomous software tool. Once the decisions about route layout were made, the actual process of route creation was fairly straightforward. The difficulty, however, was the final step: distributing the events around those routes in such a manner that accounted for all of the above constraints and considerations. The difficulty was twofold: making sure that only one target was in line-of-sight at any given time, and ensuring that each target was in sight for long enough that the participant had a reasonable chance to identify it. In order to balance those two requirements, as previously described, numerous design iterations were necessary.

Another complication was that the scenarios had to be statistically balanced between them. Every effort was made to have them be statistically similar, but even so there were minor variations between the scenarios. The four scenarios were presented in a mixed order and mixed display configurations, in order to try and equalize any scenario differences.

As part of the methodology followed by the team, a data analysis process had to be run concurrently with scenario development. The results of the data analysis process helped ensure that the scenarios were clean. An automated data analysis program was developed that could parse the scenario configuration files and determine when targets should have been visible to the participants (i.e. the targets were scripted to appear on the screen in a given camera view). Then, the custom software would analyze "line-of-sight log files" produced by the run-time-system and check that threats were only seen in the correct order, reporting an error if targets appeared out of order. This system mitigated the need for manual verification of scenario logs, improving the efficiency and accuracy of the scenario development phase.

2.4 Data Analysis

One output of the data analysis process was a file format specification for every component with logging responsibilities. Because it defined a generalized format for the logging of the experiment data, the existence of that specification enabled the team to develop the automated data analysis program at the same time as others on the team were developing the scenarios and still others were doing the necessary hardware and software integration work. As a result, a full version of the automated analysis tool was available shortly after the end of the experiment, as opposed to requiring the development team to await the complete data set before starting to create the analytic toolset. Owing to the nature of a distributed simulation engine of the type used in this experiment, there were a dozen files, each with its own set of data and time-stamps, all of which had to be merged and collated into a single timeline so that the threat reports could be properly reduced and analyzed.

The data combining process was important for reasons beyond verification of the proper execution of the simulated scenarios. It was only because the team had a basic timeline for every experimental run that they could turn to the most difficult part of data analysis. That is, the determination of the relationship between a given entity in the simulated environment and reports issued by the participant. One logging program recorded what entities were in sight at all times. This meant that when combined with the threat reports, it was possible to tell what was visible to the user during the few seconds before the report was issued. The threat report itself had the participant's thoughts as to whether the threat was an unarmed human, an armed human, or an IED, and what its location was relative to the vehicle heading.

While it would have been possible to record more precisely the axis that the participant was reporting a threat along by having them indicate via the WMI where they thought the threat was, that did not seem as valuable from the experimental perspective. That is, it was very important from a cognitive perspective to assess how effectively the participants could perform the mental rotations necessary to relate the threat location to the current vehicle heading. This is ultimately one of the most important aspects of 360° LAA, and so it was felt that letting the software handle those rotations would discard some very valuable information. Unfortunately, because of the chosen mechanism of recording threat reports (i.e. having a user interface where participants would enter particular characteristics of threats as opposed to simply utilizing the touch screen and having the participants indicate threats by touching the screen region where they appeared), determining the intended target of a threat report represented a nontrivial data reduction task. The difficulty in the process of matching up reports to events in the environment at first does not seem daunting. The team, in fact, underestimated this difficulty when developing the software to run the simulation and check for line-of-sight matching to environmental entities. Indeed, such a task is easy as long as the report is correct: if there was an armed human reported at 3 o'clock at the same time as an armed model of a human appeared to the vehicle right, then the process of attributing the threat report to its subject was simple and straightforward. However, when mistakes were made in the threat report

(i.e. an armed human at 3 o'clock was reported when both an unarmed human appeared to the right and an armed human appeared to the left) or when a distractor (non-threat) entity appeared near to the target that was supposed to be reported on, thus creating an ambiguity in terms of which entity the participant was reporting on, then the data reduction grew more difficult.

A five person working group each took a copy of the combining timeline for a single scenario-representing one-quarter of all threat reports and then spent several hours, on their own, determining relationships between each threat report and entities in the environment. The working group then reunited and spent a few more hours comparing the results of their assessments. The result of the ensuing discussion was a set of logical filters to match threat reports to entities. An algorithm was then created to apply those rules and determine a confidence level for every entity and report, and then assign high confidence matches and separating out cases where no match was possible.

The elements of the logical filter are summarized in Table 1. These elements were used to assign a quantitative score to each entity that was considered a possible subject of a given threat report; each element of the logical filter was thought to increase the probability that a given entity was a subject of a threat report and the higher the total score (sum of the elements) indicated a higher likelihood of an entity being assigned as the subject of a threat report. Once scores were assigned to each candidate entity, Boolean logic was used to compare those entities with the top two scores. In the case that there was a clear "winner", the algorithm assigned the top entity to the threat report. The "winner" was only assigned if the top entity had a score that exceeded the second highest score by a predefined margin, $\delta = 0.75/\ 6.5$ (the maximum possible score). If, however, neither of the scores exceed the minimum confidence level, $\alpha = 0.7$, or if the top score did not exceed the second highest by δ, then the algorithm reported an ambiguity that had to be resolved manually by an experimenter.

Table 1. Summary of the variables used to assign quantitative scores to each entity

Filter Variable	Description	Range of Values
IsATarget	Was the entity a scripted target?	[0, 1]
IsMoving	Was the entity moving?	[0, 0.5]
CurrentLOS	Was current line of sight (LOS) established?	[0, 1]
InClockPosition	Was the entity in the reported threat position?	[0, 1]
TypeMatch	Did the entity match the reported threat type?	[-1 : 1]
ViewingProportion	During what proportion of the previous 10 seconds was the entity visible in one of the operator's sensor views?	[0 : 1]
InRange	Was the entity within viewing range of the simulated vehicle?	[-1 : 1]

After the algorithm was developed, its results were compared against the selections made by the working group on 339 valid reports (some of the original 383 were omitted from this comparison for technical problems associated with a single subject who did not perform the reporting task consistent with instructions). This comparison showed 83% of the reports were accurately associated with entities by the algorithm, 12% were returned as ambiguous and thus required manual confirmation and 5% were judged as erroneous associations. Of the 5% erroneous associations, 2.6% were false negatives that would be caught by manual verification and another 2.3% were false positives, of which approximately 50% would have been rectified as a consequence of resolving false alarms and ambiguous results. So the ultimate rate of "undetectable error" by the algorithm, according to this assessment, was approximately 1 – 2% of all threat reports.

After all threat reports were assigned entities as just described, the metrics were examined statistically for consistent patterns and differences among experimental conditions as well as for predictive relationships (correlations) between entity characteristics and threat report answers. The statistical analyses were conducted using two different types of tests including logistic regression for the binary variable (threat detected or not) and linear mixed-model regression for assessment of the continuous variables (reaction time, threat report accuracy, subjective workload).

3. RESULTS

The data from this experiment have been previously presented[12,13] and a portion of those data are being provided here to facilitate discussion of the outcomes of the assessment methodology described above. The point deserving emphasis is how it was the rigor of the methodology that made these results accessible. The strength of the methodology both

provided enough high quality data to make counter-intuitive discoveries and the confidence that these were real effects, and not an artifact of the simulation or scenario.

Among the most surprising results from the present assessment was the simple observation that display type appeared largely unrelated to threat detection performance. Shown in Figure 4, we observe an absence of a main effect for display type on proportion of threats detected. In other words, there was little to no systematic variation in overall threat detection as a function of the display being used. Indeed, the a priori expectation was that, at minimum, access to a wider field of view through the banners would facilitate threat detection performance as compared with the more restrictive view provided by the single, narrow field of view in Display configuration A. Instead, what was seen was that the main effect appeared in terms of whether threats first appeared to the front or the rear of the vehicle regardless of the display configuration being used. There was a small interaction effect, such that the reduction in detection of threats to the rear was dependent on display configuration. In particular there was a slightly greater reduction in rear threat detection performance when participants used Display D as compared with the other displays.

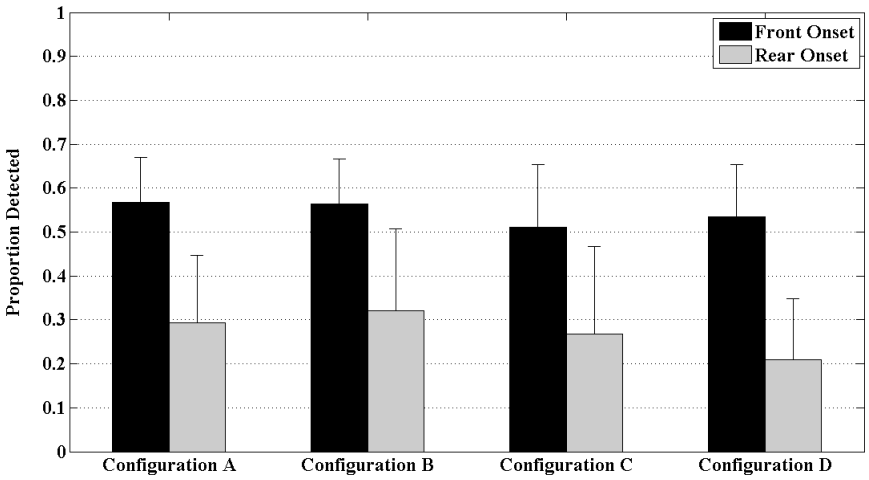

Figure 4. Display Configuration x location effect.

Figure 5 demonstrates that part of the reason that the participants performed so poorly on threats presented to the rear was because of the manner in which they used the small sensor portal that was common to all display configurations. The subjects only infrequently utilized the sensor portal to view the rear of their vehicle in all display configurations. However as discussed above, the banner displays merely provided psychological justification for this choice of strategy: they did not provide any actual performance improvement. This was shown in Figure 4, which demonstrates that Display configuration D had the worst performance for rear onset targets (the Display configuration x location interaction effect).

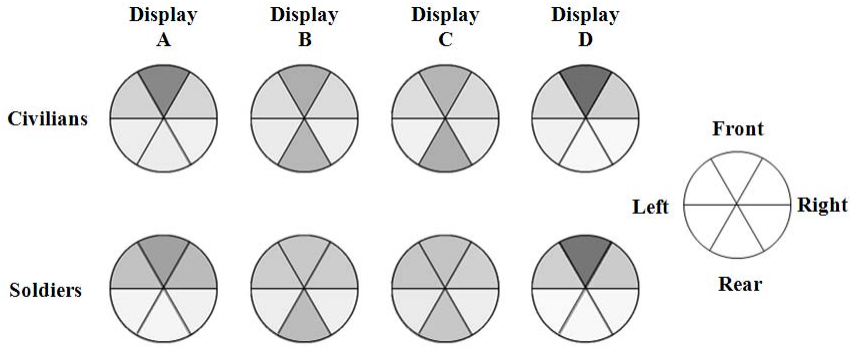

Figure 5. Proportional usage of the small sensor portal as a function of sensor view direction and configuration; darker shading indicates greater proportion of time spent viewing that vehicle-relative direction.

As is apparent in Figure 5, the participants tended to focus disproportionately on the front views during the experimental trials and showed the heaviest bias towards viewing the front central sensor view, a behavior which was particularly pronounced while using Configuration D more than in any other configuration. This is also confirmed from looking at Table 2, which shows the number of view changes in an average data collection, showing that in Configuration D the participants engaged their central sensor portal much less than in other configurations.

Table 2. Sensor View Changes as a Function of Display Configuration

Display Configuration	View Changes per mission	View Changes Per Minute
A	381.5	28.2
B	171.4	12.6
C	189.2	13.8
D	131.0	9.6

The limited eye-tracking data, when analyzed, not only supported, but also supplemented this conclusion: the 7 participants for whom high quality eye tracking data existed spent 3% of the time looking at the bottom banner. It appears that the bottom banner provided a cognitive effect that could be labeled as "false confidence": the participants believed that they had coverage of the rear sector provided from the banner, but they rarely actually looked at it long enough to gain actual local area awareness. Indeed, the limitations of subjective feedback became apparent in light of these data. The team had hypothesized that banners would yield significantly improved performance over the single sensor portal, and in particular that Configuration D would provide superior threat detection performance in the rear sector; the subjective self-reported data agreed. That is, banners were strongly preferred and Display configuration D was the favorite of the participants in exit interviews. That these preferences were not supported by the actual quantitative data from the experimental runs must be seen as a triumph of the strong, structured methodology that underpinned the assessment.

The study did provide confirmation of one hypothesis that the team had suspected. Specifically, threat detection performance fell off dramatically as a function of threat range from the vehicle. This was largely because even in the highest resolution configuration, human-sized targets would be very small at over 100 meters without the capability to zoom. Perhaps more interesting, however, was that this reduction in threat detection performance as a function of range differed dependent on the type of target that was being observed. Shown in Figure 6, it can be seen that (a) IEDs were always detected at a high rate, though always at short ranges and (b) performance in detecting Armed Humans was considerably more variable than performance in detecting Unarmed Humans, particularly at short ranges.

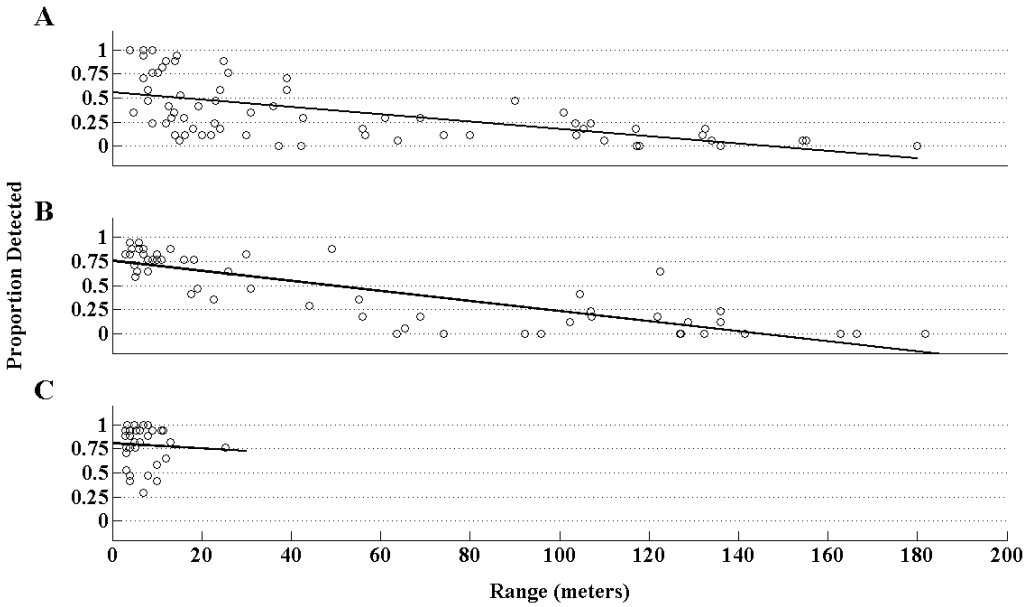

Figure 2. Average threat detection performance as a function of range and target type. Target Type A: Armed Humans Target Type B: Unarmed Humans Target Type C: IED

The observation that threat detection performance was superior when looking at Unarmed Humans under any conditions appears somewhat counterintuitive at first. That is, when discussing the constraints on scenario design, the experimental team considered the brandishing of a weapon to be a relatively obvious cue whereas it was thought that the fact that Unarmed Humans were only distinguished as "report-worthy" by their behavior was considered to be more subtle. Yet, again we have a situation where the quantitative data reflected a different story. In the interpretation of the experimental team, this difference was likely due to a difference in target salience. That is, Unarmed Humans were often only considered to be threatening if they were actually engaging (physically moving towards) the vehicle whereas Armed Humans were often standing still or even walking away from the vehicle. Thus again, we have a situation where careful scenario design taking into account and incorporating a range of behaviors from the military context allowed for a quantitative analysis that revealed counterintuitive results that may have been overlooked in more standard human factors and engineering assessments of systems usability and utility (i.e. range performance results were as expected on a course level, but showed interesting results in a more detailed analysis).

Finally, because it was a central consideration in scenario design, we discuss the reaction time results. As with all of the other variables, reaction times revealed a complex pattern. However, in general the reaction time data suggested a single general conclusion – as participants had more time available to view a given target, they took advantage of that time to confirm its identity before filling out a threat report. This general pattern is revealed in the data shown in Figure 7, which represents a relatively strong correlation between amount of time available for viewing a target and the reaction time. Overall, these data were interpreted as reflective of a general cognitive strategy wherein the participants were prioritizing response accuracy over response speed.

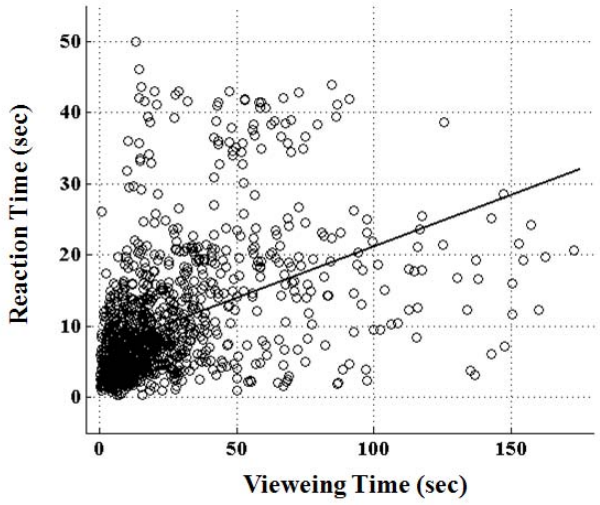

Figure 7. Reaction time (time between an appearance of a target and a participant's initiation of a threat report regarding that target) as a function of viewing time (amount of time the target was viewable).

4. DISCUSSION

The present paper provides a detailed discussion regarding the structure and execution of an evaluation of a display system designed to enhance Soldier situational awareness for local area security tasks within military-relevant environments. The intent of this discussion was to convey the importance and value of executing a thoughtful, well-researched and yet complex experimental design. Indeed, the process of creating an evaluation of a 360° LAA system is extraordinarily tricky. Given modern computational technology, the task of having entities move about a virtual environment is generally straightforward (if it is not then the simulation engine interface needs improvement). However, when precise sets of events with particular characteristics are needed (especially when those events must be reactive to the operator and/or his vehicle), then scenario design becomes elevated in its status with respect to experimental design. For the present paper, it was shown that to get the timing right, and to create enough entities moving about the simulated terrain in a controlled fashion for the test to be a realistic evaluation of using the display in a operationally relevant situation is a difficult and yet valuable process. The experimental team was required to consider: (a) the intent of the system design – to be able to provide access to imagery of the environment in a 360° array surrounding a moving tactical vehicle, (b) the nature of the environment within which the system would be deployed and (c) the nature of the human responses and behavioral tendencies with such a system and interface within the context of deployment. Ignoring any of these aspects while creating the assessment ultimately results in risking information loss. In the context of military systems, this may mean sacrificing performance – which could in turn compromise the lethality and survivability of American Soldiers.

REFERENCES

[1] Department of the Army, "Army Modernization", http://www.bctmod.army.mil/ (2011).
[2] Department of the Army, "2010 Army Modernization Strategy", https://www.g8.army.mil/pdf/AMS2010_hq.pdf (2010).
[3] Demming, C., "TARDEC ATOs Aim to Reduce Soldier Workloads", Army AL & T Online January(2009).
[4] Parasuraman, R., Barnes, M., and Cosenzo, K., "Adaptive automation for human-robot teaming in future command and control systems", The International C2 Journal 1(2), 43-68 (2007).
[5] Shoemaker, C. M. and Bornstein, J. A.,"The demo III UGV program: A testbed for autonomous navigation research", Proceedings of IEEE Intelligent Control (ISIC), 1998.,(1998).
[6] Goodwin, N. C., "Functionality and usability", Communications of the ACM 30(3), 229-233 (1987).
[7] Grudin, J., "Utility and usability: Research issues and development contexts", Interacting with Computers 4(2), 209-217 (1992).

[8] Scholtz, J.,"Beyond usability: Evaluation aspects of visual analytic environments", Proceedings of the 2006 IEEE Symposium on Visual Analytics Science and Technology,(2006).
[9] Shackel, B., "Usability - Context, framework, definition, design and evaluation", in B. Shackel and S. Richardson (Eds.) [Human factors for informatics usability], Cambridge University Press, Cambridge, UK, 21-37 (1991).
[10] Maguire, M., "Context of use within usability studies", International Journal of Human-Computer Studies 55, 453-483 (2001).
[11] Hart, S. G. and Staveland, L. E., "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research", in P. A. Hancock and N. Meshkati (Eds.) [Human Mental Workload], North Holland Press, Amsterdamn, 239-250 (1988).
[12] Gordon, S., Cosenzo, K., and Brumm, B.,"Concurrent Assessment of 360° Local Area Awareness using Physiological and Behavioral Metrics", 27th Army Science Conference,(2010).
[13] Metcalfe, J. S., Cosenzo, K. A., Johnson, T., Brumm, B., Manteuffel, C., Evans, A. W., and Tierney, T.,"Human dimension challenges to the maintenance of local area awareness using a 360° indirect-vision system", 2010 NDIA Ground Vehicle Systems Engineering and Technology Symposium: Modeling and Simulation, Testing and Validation Mini-Symposium,(2010).