

Use of Real-Time Probes for Measuring Situation Awareness

Debra G. Jones and Mica R. Endsley

*SA Technologies, Inc.
Marietta, Georgia*

This study examined the validity of real-time probes as measures of situation awareness (SA). Real-time probes are verbal queries (derived from an SA requirements analysis) posed to the operator concurrent with operations. Mixed results were obtained. A weak but significant correlation was found between real-time probes (both accuracy and latency measures) and Situation Awareness Global Assessment Technique queries, indicating that real-time probes were measuring some facet of SA. However, correlations with workload were also found, and this correlation needs to be investigated further. Although real-time probes show promise, more research is needed to assess the utility of real-time probes as a metric of SA.

With the ever-increasing complexity of systems, operators can quickly become cognitively overtaxed and unable to adequately process the large amounts of data with which they must contend. Effectively processing data that are uncertain and constantly changing is essential to understanding the current state of the environment and to successfully predicting the evolution of the environment over time. The term *situation awareness* (SA) refers to the ability of people to develop an internal representation of the current state of the environment and to project likely future states of the environment. Formally, SA can be defined as “the perception of the elements within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future” (Endsley, 1995a, p. 36). To make informed decisions, the operator must be cognizant of all the relevant elements of the environment, what these elements mean, and how these elements will affect the environment over time. Thus, maintaining a high level of SA is essential for effective

decision making and designing systems to assist operators in developing and maintaining SA facilitates decision making.

To evaluate whether a system adequately supports operator SA, a method must be developed to measure the SA afforded by the system. Unfortunately, some difficulties may be encountered when trying to measure this complex construct. Although both subjective and objective metrics have been developed to measure SA, even the most successful measures fall short of being able to assess SA during actual operational tasks.

Several subjective metrics of SA have been developed that could be employed to assess SA for operational tasks. For example, the Situational Awareness Rating Technique (SART) can be administered to participants at the conclusion of an operation (Taylor, 1990). This measure consists of a 10-dimensional bipolar scale on which participants can subjectively rate their SA. The ratings provided by the participants on each of the 10 items are then combined to form a rating for each of three major factors: supply of attention, demand for attention, and understanding. Although this measure effectively provides information regarding a participants' confidence in their SA, it does not provide an objective measure of SA. It can be influenced by memory decay (as it is taken at the end of an event) and by performance outcome (i.e., a person who successfully performs the task may rate SA higher based on the positive outcome of the event). More important, people do not know what they do not know and thus may be poor at accurately assessing their own SA. In addition, subjective measures rarely provide detailed diagnostic information regarding the system.

Objective measures avoid the pitfalls associated with subjective measures, but these types of measures are often more difficult to implement in real-time activities. For example, the Situation Awareness Global Assessment Technique (SAGAT) provides an objective measure of SA (Endsley, 1988, 1995b); however, this method is designed to be used in a simulation environment and is more difficult to use in many real-time operational environments. With this technique, the simulation is frozen at randomly selected times, the simulation is suspended, and the system displays are blanked while the operator quickly answers questions about his or her current understanding of the situation. Operator perceptions are then compared to the real situation (based on information drawn from the computer or from subject matter experts (SMEs) who answer the SAGAT queries while looking at the displays). Comparing the data in this manner provides an objective, unbiased assessment of SA.

The SAGAT technique has been shown so far to have a high degree of validity for measuring SA. SAGAT has been shown to have predictive validity, with SAGAT scores indicative of pilot performance in a combat simulation (Endsley, 1990a). Content validity was also established, showing the queries used to be relevant to SA in a fighter aircraft domain (Endsley, 1990b). Empirical validity has been demonstrated through several studies showing that a temporary freeze

in the simulation to collect SAGAT data did not impact performance and that such data could be collected for up to 5 or 6 min during a freeze without running into memory decay problems (Endsley, 1990b, 1995b). A certain degree of measurement reliability has been demonstrated in a study that found high reliability of SAGAT scores for individuals who participated in two sets of simulation trials (Endsley & Bolstad, 1994). However, as SAGAT is difficult to implement in many real-time operational systems (e.g., the flight environment) due to the problem of freezing the action, another approach may be needed to evaluate SA in these types of environments.

Thus, a gap exists between effectively measuring SA in operational environments and the availability of a metric that can afford a valid measure of SA during real-world operational tasks. Such a measure should neither influence the operator's SA by its presence nor distract the operator from essential tasks, thereby compromising safety and influencing ongoing task performance.

In an effort to meet this need, Durso et al. (1998) used the Situation-Present Assessment Method to assess the SA of air traffic controllers. With this method, the participants were asked queries regarding the situation at periodic intervals, but the displays remained in full view of the participants. Response latency was measured to provide an indicant of the quality of the participants' SA. Theoretically, participants who had good SA would answer more quickly because they would know where in the environment to look for a particular piece of information and thus could answer the question faster. This study found that controller reaction time to the probes about the current status of events in the simulation (Level 1 SA) were correlated with the SME's subjective ratings of controller performance ($R^2 = .53$). Probes about future (Level 3 SA) were correlated at a much lower level, with a measure of how many actions the controllers still had left to complete at the end of the simulation ($R^2 = .12$). Although this study suggests that these probes have promise as a measure of SA, a detailed assessment of the validity of the probes was not conducted.

In this article we describe efforts to further investigate the concept of measuring SA by querying the participant concurrently with ongoing operations. The methodology was derived from SAGAT and allows operators to be verbally queried concurrently with real-time activities. Like SAGAT, these queries (called *real-time probes*) are developed from an extensive SA requirements analysis that delineates the operator's goals and the associated SA requirements. The probes are designed to specifically query an aspect of SA needed for task performance. Unlike SAGAT (in which operations are suspended and operators are presented with a series of questions) however, the operator is periodically presented a single verbal query and is required to verbally respond.

A number of issues need to be addressed to ensure the validity of real-time probes as a measure of SA. Because operators may simply look for information to answer each probe, accuracy might provide very little information about operator

SA. Reaction time to each probe may theoretically correspond to the degree to which the requested information is already known (e.g., in working memory) or understood (e.g., the operator is fully aware of the situation and knows where to quickly find the information). We examine these response measures (accuracy and response time) to the real-time probes to determine their validity as indicants of SA in real-time, dynamic environments.

In addition, the intrusiveness of probes posed concurrently with ongoing operations is a concern. Such probes may either be distracting, providing additional loading that would take away from primary task performance, or provide help, cueing operators to attend to information they might otherwise have missed.

Finally, the potential for such a measure to reflect constructs other than SA must be considered. Real-time probes are actually fairly similar to secondary workload probes. Reaction time to these probes may be more indicative of the level of workload the operators are under than their SA. Thus, the possibility that real-time probes provide a measure of workload rather than SA must be examined.

To evaluate the validity of real-time probe measures for assessing operator SA, a series of studies were conducted.

STUDY 1

This study was designed to evaluate two air traffic control (ATC) displays at the Federal Aviation Administration William J. Hughes Technical Center Research Development and Human Factors Laboratory (first reported in Endsley, Sollenberger, Nakata, & Stein, 2000). The simulation facility at this laboratory supports high-fidelity simulations and incorporates controller workstations representative of those found in the field—complete with radar displays, keyboards, and flight strip bays. The test was conducted using the ATCoach™ Version 7.0 (1996) simulation system (UFA Inc., Woburn, MA) operating on a Sun® workstation. ATCoach provides a realistic, high-fidelity simulation of a controller's workstation. It includes a 20 × 20 in. high-resolution color radar display monitor (2000 × 2000 pixels), a three-button trackball, and keyboard. A flight strip bay with printed, standard configuration flight strips provides information for each aircraft in the simulation. Furthermore, the facility is equipped with a voice communication system that links the controllers with confederate pilots in a remote room. These confederates are responsible for moving the aircraft in the simulation in accord with controller commands as well as a priori instructions. Although the main objective of the study was to assess display types, the study also necessitated the collecting of several measures of SA (including real-time probes and SAGAT) as well as measures of workload and performance. The results for the real-time probes were compared to these other concurrent measures to provide an initial assessment of their validity.

Method

Ten 45-min scenarios were created for this study. Two training and three test scenarios were developed for each of two display formats. The scenarios all involved similar traffic density and complexity; approximately 32 aircraft were involved over the course of the scenarios, but the number of aircraft on the screen at any one time varied significantly over the course of the trial. All trials took place in a generic airspace high-altitude sector (Genera Sector) that was specifically designed for human factors research testing (Gutman, Stein, & Gromelski, 1995).

In all of the trials, the aircraft operated under a free flight self-separation operational concept—that is, the pilots were responsible for maintaining separation rather than relying on the controller for guidance. The controllers were given instructions regarding the conditions that warranted the issuing of warnings and advisories for aircraft that did not appear to be maintaining adequate separation. When issued advisories, the pilots had the discretion either to follow the command or to take other actions. The scenarios were scripted so that, in general, the aircraft maintained at least the required minimum separation from other aircraft. In each scenario, however, a few problems were added in which the aircraft did not maintain the required separation, thereby requiring controller intervention to ensure separation.

Participants

Ten experienced air traffic controllers (M experience = 12.1 years at full performance level, range = 3.9 to 28.6 years) from five different en route ATC centers participated in this study. All of the controllers were current and had at least 16 hr controlling traffic in the preceding months. Participation in the study was voluntary. Additional participants included the confederate pilots in the remote room who followed a scripted plan of action and experienced controllers who served as SMEs. The SMEs had a variety of responsibilities: They provided a subjective rating of each participant's performance at the conclusion of each trial, provided an overall rating describing how well the controller managed traffic on a 10-point scale, provided a subjective assessment of workload, and completed the data collection forms for SAGAT and the real-time probes to enable accurate scoring of the participant's responses.

Data Collection

Measures collected during this study included three SA measures (real-time probes, SAGAT, and SART), two workload measures (the Air Traffic Workload

Input Technique [ATWIT] and the NASA Task Load Index [NASA-TLX]), and numerous performance measures.

1. SA measures

- **Real-time probes:** The real-time probes are questions posed to the operator concurrent with the simulation. In this study, the probes were delivered over the participants' headsets by an experimenter posing as an adjacent sector's controller. Eight probes, corresponding in content to eight of the SAGAT queries, were administered (see Table 1). These probes occurred either 1 min before or 1 min after the ATWIT workload measurement. The probes were linked to the ATWIT measurements to allow for comparison between probe reaction time and the ATWIT measure. The eight probes were administered in a random order, and response time and accuracy were recorded.

- **SAGAT:** Four freezes were inserted at randomly selected times during the trials, and responses to the entire battery of SAGAT queries (16 queries; see Table 2) were collected at each stop via the HyperCard program on a Macintosh® computer placed adjacent to the controller's station. Each freeze lasted approximately 4 min.

- **SART:** SA was also measured using the SART (Taylor, 1990). This measure, a 10-dimensional bipolar scale on which participants subjectively rate their SA, was completed by participants at the end of each trial using HyperCard on a Macintosh computer. The ratings provided by the participants on these 10 items were combined to form a rating for each of the scale's three major factors (supply of attention, demand for attention, and understanding) as well as an overall score.

TABLE 1
Study 1: Real-Time Probes

SA Level 1—Perception of traffic situation

- . What is the current heading for aircraft X?
- . What is the current flight level for aircraft X?
- . Climbing, descending, or level: Which is correct for aircraft X?
- . Turning right, turning left, or on course: Which is correct for aircraft X?

SA Level 2/3—Comprehension and projection of traffic situation

- . Which aircraft have lost or will lose separation if they stay on their current (intended) course?
 - . Which aircraft will be affected by weather within the next 5 minutes, unless an action is taken to avoid it?
 - . Which aircraft must be handed off within the next 3 minutes?
 - . What is the next sector for aircraft X?
-

Note. SA = situation awareness.

TABLE 2
Study 1: Situation Awareness Global Assessment Technique Queries

-
1. Enter the location of all aircraft (on the provided sector map).
 - Aircraft in track control
 - Other aircraft in sector
 - Aircraft will be in track control in next 2 minutes
 2. Enter aircraft callsign (for aircraft highlighted of those entered in Query 1).
 3. Enter aircraft altitude (for aircraft highlighted of those entered in Query 1)
 4. Enter aircraft groundspeed (for aircraft highlighted of those entered in Query 1)
 5. Enter aircraft heading (for aircraft highlighted of those entered in query 1)
 6. Enter aircraft's next sector (for aircraft highlighted of those entered in Query 1).
 - A B C D E G
 7. Enter aircraft's current direction of change in each column (for aircraft highlighted of those entered in Query 1).

<u>Altitude change</u>	<u>Turn</u>
climbing	right turn
descending	left turn
level	straight
 8. Enter the aircraft type (for aircraft highlighted of those entered in Query 1).
 9. Which pairs of aircraft have lost or will lose separation if they stay on their current (intended) courses?
 10. Which aircraft have been issued advisories for situations which have not been resolved?
 11. Did the aircraft receive its advisory correctly (for each of those entered in Query 1)?
 12. Which aircraft are currently conforming to their advisories (for each of those entered in Query 1)?
 13. Which aircraft must be handed off to another sector/facility within the next 2 minutes?
 14. Enter the aircraft which are not in communication with you.
 15. Enter the aircraft that will violate special airspace separation standards if they stay on their current (intended) paths.
 16. Which aircraft are weather currently an impact on or will be an impact on in the next 5 minutes along their current course?
-

2. Workload: Three subjective measures of workload were collected during the study. First, ATWIT probed controllers to rate their workload on a 10-point scale every 5 min (Stein, 1985). Second, the participants completed the NASA-TLX workload survey (Hart & Staveland, 1988) at the end of each trial via HyperCard on a Macintosh computer. Finally, both the participants and the SMEs provided a subjective assessment of workload at the end of each trial for the participants by designating the participants' level of workload on a 10-point scale.

3. Performance measures: Both subjective and objective performance measures were collected during the study. The subjective measures included a subjective rating of each participant's performance using a modified version of the Observer Checklist (Sollenberger, Stein, & Gromelski, 1997). The SMEs completed these rating forms at the end of each trial. In addition, at the end of each trial the SMEs provided an overall rating (on a 10-point scale) describing how well the

participant managed traffic. Finally, a series of objective performance measures were collected during each trial (Table 3).

Procedures

Each of the participants took part in six test trials, three in each display condition. The six scenarios were counterbalanced across trials and participants. The participant completed the study over the course of 3 consecutive days. Prior to the start of the study, the participants received instructions regarding real-time probes, SAGAT, SART, ATWIT, and NASA-TLX. The participants also received 1 hr of familiarization training on the ATCoach simulator, the Genera Sector, and the operational concept.

Prior to each experimental display condition, the participants received two training trials, which included two practice SAGAT stops and four real-time probes. Following the practice trials for each display condition, the participant completed the three test trials for that condition. During each testing trial, four SAGAT freezes and eight real-time probes were administered. In addition, an audio tone sounded every 5 min, prompting the participant to fill out the ATWIT scale. At the end of each trial, the participants completed the SART and NASA-TLX surveys as well as a postscenario questionnaire. At the same time, the SMEs filled out the observation checklist form rating the participant's performance.

TABLE 3
Objective Performance Measures

Safety of flight
Number of enroute conflicts
Duration of enroute conflicts
Number of between sector conflicts
Duration of between sector conflicts
Efficiency
Number of flights handled
Duration of flights handled
Distance flown in sector
Number of completed flights
Cumulative average aircraft density
Control strategy
Number of changes in altitude/aircraft handled
Number of changes in speed/aircraft handled
Number of changes in heading/aircraft handled
Communication and taskload
Number of controller transmissions
Duration of controller transmissions

Results

An analysis of the displays used in the study can be found in Endsley et al. (2000). Only those factors relating to the viability of real-time probes as a measure of SA are considered here. Two aspects of the real-time probe must be examined when considering its potential use as a metric of SA. First, the measure must possess sensitivity equivalent to other measures. Second, the real-time probe must be shown to measure SA rather than some other construct, thereby demonstrating the validity of the metric. The accuracy of the real-time probe was high (95%); however, because the displays remained in view, this fact does not necessarily indicate that the participants had a high level of SA. Even without a high level of SA, the participants could have searched the display for the appropriate answer. Theoretically, shorter response times may be indicative of the participant's being aware of the information and not having to search the display for the information. Thus, shorter responses times are indicative of a higher level of SA. Consequently, only response time (to correct answers) was used in these analyses.

Sensitivity

To fully assess the real-time probe, a comparison must be made between the real-time probe and existing measures of SA. The sensitivity of real-time probes should be comparable to that of existing metrics. Two display conditions were present in this study, thereby providing the opportunity to examine the sensitivity of the various measures to display manipulations. The various measures (real-time probes, SAGAT, SART, ATWIT, NASA-TLX, and performance measures) were analyzed using analyses of variance (ANOVAs) to determine which demonstrated such sensitivity. Only four elements showed sensitivity to display manipulations:

- One element of the SAGAT measure: knowledge of aircraft conformance to advisories, $F(1, 9) = 6.961, p = .027$.
- One element of NASA-TLX measure: mental workload component, $F(1, 9) = 23.070, p = .001$.
- One objective performance measure: efficiency (number of completed flights), $F(1, 9) = 8.620, p = .017$.
- One subjective performance measure: maintaining attention and SA (correcting own errors in a timely manner), $F(1, 9) = 5.847, p = .039$.

ATWIT, SART, and the real-time probes did not differentiate between display conditions ($\alpha = .05$). Because many of the measures were not sensitive to the display manipulation, the lack of sensitivity of the real-time probe cannot be taken as a strong indication of this measure's inability to differentiate between

conditions. Thus, the sensitivity of the real-time probe could not be adequately assessed from these data.

Validity

A major concern in assessing the utility of real-time probes is whether the probes actually measure SA. Theoretically, reaction time to real-time probes provides an indication of the degree to which needed information is actively present in working memory, thereby providing an index of SA. In reality, another explanation is possible: reaction time to the real-time probe may provide an indication of the operator's spare mental capacity and is thereby functioning as a secondary workload task. To discern which of these two explanations is most likely, a stepwise regression was performed to compare the real-time probe data to SAGAT and to ATWIT. No relation was found between participant reaction time to the real-time probes and their mean accuracy on the corresponding SAGAT query ($p > .05$), thereby failing to give support to the assertion that real-time probes measure SA. Furthermore, a weak correlation was found between probe reaction time and reaction time to the ATWIT tone, $F(2, 477) = 3.48$, $p = .032$, $R^2 = .014$, indicating that the real-time probe may instead be providing a measure of workload.

In another study examining online probes, Durso et al. (1998) examined the ability of the probes to predict performance in an ATC task. Two categories of probes were created: probes about the present situation and probes about the future situation. These measures were found to be somewhat predictive of performance for the tasks in the study. To replicate this approach, the real-time probes and SAGAT queries collected in this study were combined in similar categories: Level 1 SA probes equated with situation present probes and Level 2/3 SA probes equated with the future situation. These combined measures (as well as an SME SA rating and the three major SART components—supply, demand, and understanding) were then compared to performance measures to examine their predictive utility. A stepwise regression was performed comparing the SA measures to five performance measures: SMEs' overall rating of performance, the mean of all the subjective ratings of controller performance, number of flights completed (the only measure sensitive to display manipulation), the number of conflicts in the trial, and the duration of conflicts. The statistically significant results for these measures are summarized in Table 4.

Few of the SA measures showed significant predictive ability for the factors considered. Only two significant relationships were found: SA Level 2/3 probe mean reaction time correlated with the number of conflicts and with the duration of conflicts. Surprisingly, the relationship between Level 2/3 probe mean reaction time and the number and duration of conflicts was in the opposite direction from expected; slower reaction time for the comprehension and projection

TABLE 4
Summary of Regression Models for Predictability of Performance From SA Measures

<i>Model</i>	<i>F Value</i>	<i>p</i>	<i>Adjusted R²</i>
SME overall performance rating			
SME SA Rating	$F(1, 52) = 152.00$	< .001	.727
SME mean performance rating			
SME SA Rating	$F(1, 58) = 287.45$	< .001	.832
SART—Understanding	$F(1, 57) = 7.93$.007	.122
Overall model	$F(2, 56) = 147.26$	< .001	.840
No. of flights completed			
None			
No. of conflicts			
SA Level 2/3 probe RT	$F(1, 57) = 6.14$.016	.097
SME SA rating	$F(1, 57) = 5.49$.023	.088
Overall model	$F(2, 56) = 5.20$.008	.157
Duration of conflicts			
SA Level 2/3 probe RT	$F(1, 57) = 5.23$.026	.084
SART—Demand	$F(1, 56) = 4.26$.044	.071
SME SA rating	$F(1, 57) = 5.23$.023	.084
Overall model	$F(3, 54) = 6.41$.001	.263

Note. SA = situation awareness; SME = subject matter expert; SART = Situational Awareness Rating Technique; RT = reaction time.

real-time probes was equated with fewer ATC conflicts in the simulation. Similar results were found for the conflict duration. The reason for these findings is unclear, but it may have stemmed from the controllers feeling rushed when more conflicts were ongoing and therefore responding more quickly. Alternatively, the controllers who took longer to answer the questions may have been employing a different information-gathering technique in which they spent more time assessing the situation, thereby allowing them to react more decisively and resolve the conflict quicker. Thus, the increased reaction time may be reflective of increased information-gathering strategies. In either case, this finding causes concern as to whether operator reaction time to the real-time probes is an appropriate measure of SA.

Neither of the combined SAGAT measures showed any relationship with the five performance measures. However, this result was not surprising, as responses to difficult SAGAT queries tend to be independent of each other (Endsley, 1990a). The combined measure was used here to provide a comparison with the results found by Durso et al. (1998). When examined individually, the SAGAT queries did show significant predictive ability not only for the five measures just considered but also with other performance measures (Table 5).

TABLE 5
Summary of Regression Models for Predictability of Performance From SAGAT

<i>Model</i>	<i>F Value</i>	<i>p</i>	<i>Adjusted R²</i>
SME overall performance rating	$F(1, 57) = 4.79$.033	.078
Aircraft with advisories			
SME mean performance rating		<i>ns</i>	
Number of conflicts	$F(5, 51) = 4.10$.003	.287
Vertical change			
Type			
Level of control			
Aircraft separation			
Advisory reception			
Duration of conflicts	$F(4, 55) = 3.03$.025	.180
Vertical change			
Level of control			
Aircraft separation			
Aircraft with advisories			
Number of flights handled		<i>ns</i>	
Duration of flights handled	$F(3, 52) = 5.29$.003	.234
Location of sector aircraft			
Speed			
Special airspace separation			
Distance Flown	$F(3, 49) = 5.22$.003	.242
Location of sector aircraft			
Advisory reception			
Special airspace separation			
Number of completed flight		<i>ns</i>	
Cumulative average aircraft density	$F(3, 55) = 5.11$.003	.218
Altitude			
Aircraft separation			
Aircraft in communication			
Number of heading changes	$F(5, 51) = 2.64$.035	.204
Callsign number			
Altitude			
Vertical change			
Number of altitude changes	$F(3, 53) = 4.12$.011	.189
Callsign number			
Advisory conformance			
Weather impact			
Number of transmissions	$F(3, 49) = 4.025$.012	.198
Next sector			
Advisory reception			
Special airspace separation			
Duration of transmissions	$F(4, 48) = 6.187$	< .001	.340
Location of sector aircraft			
Location of all aircraft			
Advisory reception			
Special airspace separation			

Note. SAGAT = Situational Awareness Global Assessment Technique; SME = subject matter expert.

Summary of Study 1

This study did not present any evidence supporting real-time probes as a valid measure of SA. However, several factors may have affected this result. First, this study was not specifically designed to evaluate real-time probes; the analysis was a secondary objective. Next, the study's premise, a self-separation concept for air traffic controllers, was a hypothetical environment and was viewed as unrealistic by the air traffic controllers. Thus, the artificiality of the environment may have negatively impacted SA in general, making the assessment of the adequacy of real-time probes questionable. Although one concern regarding the use of real-time probes was alleviated by the controllers rating them as having low intrusiveness (M score = 2.2, on a scale ranging from 1 [*low*] to 10 [*high*]), the study did highlight some concerns associated with the use of real-time probes as a metric of SA. First, real-time probes did correlate weakly with the ATWIT workload measure, suggesting that the measure was influenced by workload. Second, the inverse correlation between the real-time probe reaction time and the number and duration of ATC conflicts (a primary performance measure) raises questions concerning the premise of using real-time probe reaction time to measure SA. That is, if reaction time to the probe provides an indication of the degree to which needed information is held in working memory, why would slower reaction time correlate with fewer conflicts? Finally, unlike SAGAT, the real-time probes showed no predictive ability when compared to select performance measures.

STUDY 2

A second effort to examine the functionality of real-time probes as a measure of SA was conducted to address some of the issues raised by the first study. This study utilized the Regional/Sector Air Operations Center (R/SAOC) training simulators at Tyndall Air Force Base. The R/SAOC team is responsible for maintaining air sovereignty in peacetime and defending North America during wartime. Their mission is accomplished by identifying all aircraft entering U.S. airspace and intercepting aircraft that either cannot be identified or are identified as a threat. Data were collected from four positions within the R/SAOC team: system surveillance technician, identification technician, and the weapons team (a weapons director and a weapons director technician). This study was designed to accomplish two things: (a) to provide a baseline measure of SA, workload, and performance on the current R/SAOC system; and (b) to examine the intercorrelation between measures of SA to assess the validity of real-time probes as an SA metric.

Method

Two 60-min scenarios (one peacetime and one wartime) that incorporated all aspects of the air-sovereignty team were developed. Data were collected from three stations within the team: surveillance, identification, and weapons team (Weapons Director and Weapons Direct Technician).

Participants

Experienced R/SAOC personnel voluntarily participated in the study, which occurred during normal duty hours. Five teams participated in the study, and each team participated in both the peace and war scenarios. Each team comprised four operators: a surveillance technician (M experience = 3.6 years, range = 1 to 8 years), an identification technician (M experience = 3.5 years, range = 1 to 10 years), and a two-member weapons team (M experience = 3.7 years, range = .5 to 8 years). Other members of the R/SAOC participated in the study as confederates to (a) administer the simulated scenarios and ensure that scenarios were as controlled as possible and (b) provide correct answers for the various queries.

Data Collection

Measures collected during this study included three SA measures (real-time probes, SAGAT, and SART), two workload measures (NASA-TLX and secondary workload), and several performance measures.

1. SA measures

- Real-time probes: Sixteen probes were presented to the participants at each station during the study. These probes were position specific and were similar in content to the SAGAT queries. The number of probe repetitions depended on the number of relevant queries for each position (see Table 6). The probes were posed to the operator one at a time at randomly selected intervals; furthermore, they were not administered while the operator was verbally communicating. As in the first study, the probes were administered verbally during the simulation, and the display remained in the participant's view. Response accuracy and response time were collected. SMEs familiar with the scenario were listening to the queries and provided an answer key to allow for later scoring of the accuracy of the participant's responses to the probes.

- SAGAT: Six SAGAT freezes were inserted at randomly selected times during the scenario, and responses were collected via pencil and paper. The SAGAT questions for each position are shown in Table 7. The SMEs provided the correct answers to each query at each stop. Again, response accuracy provided the measure for SAGAT.

TABLE 6
Study 2: Real-Time Probes

	<i>Peace</i>	<i>War</i>
Surveillance		
Which tracks have emergencies?	✓	✓
Which targets need symbology, are special tracks, or unknowns?	✓	✓
Which noninitiated tracks, special tracks, or unknowns have changed heading?	✓	✓
Which noninitiated tracks, special tracks, or unknowns have changed code/mode?	✓	✓
Identification		
Which targets are pending?	✓	✓
If you have a target that needs ID, is it fast or slow?		
If you have a target that needs ID, which agency is responsible for its airspace?		
If you have a target that needs ID, how much time is remaining for ID?	✓	✓
If you have a target that needs ID, what is its altitude?	✓	
If you have a target that needs ID, what is its code?	✓	✓
If you have a target that needs ID, what is its direction of flight?	✓	
If you have a target that needs ID, which Agency is responsible for its airspace?	✓	
How many unknowns?		✓
How many targets need ID?		✓
If you have unknowns, how many riders have been initiated for each?		✓
Weapons team		
Bearing and range		
Bullseye to tanker (D)		✓
Fighter to target (B)	✓	✓
Bullseye to western chaff dropper (D)		✓
Western fighter to E3 (D)		✓
Bullseye to E-3 (T)		✓
Fighter to intercept (B)	✓	
Fighter		
Callsign (B)	✓	✓
Altitude (B)	✓	✓
Current target (B)		✓
Targets destroyed (B)		✓
Speed (B)	✓	✓
Missiles expended		✓
Playtime (B)	✓	✓
Committed against (D)		✓
Frequency (T)	✓	✓
Time to intercept (B)	✓	✓
Intercept point (T)	✓	✓
Intercept over water (T)	✓	
Heading (T)	✓	
Mission type (B)	✓	

(continued)

TABLE 6 Continued

	<i>Peace</i>	<i>War</i>
Target		
Wartime		
Heading (T)	✓	✓
Frequency (D)	✓	
Altitude (D)	✓	
Speed (B)	✓	
Suspect by customs (B)		
Other		
Specials (B)		✓
Unknowns/Fakers (B)		✓
Distance to inner ADIZ (B)	✓	
Weather an impact (B)	✓	

Note. ID = identification; D = Weapons Director; T = Weapons Director Technician; B = both; ADIZ = Air Defense Identification Zone.

- SART: Participants completed the SART survey at the end of each trial. In addition, SMEs were asked to fill out an Observer–SART rating for the participants at the end of each trial.

2. Workload: Workload was measured in two ways during this study. First, a secondary task was added to the scenario. For the surveillance and identification technicians, this task involved a verbal response (“roger”) to a verbal cue (“acknowledge”). For the weapons team, the weapons director was asked to respond via a switch to a light that appeared on the display console. Response time to the secondary task was collected for comparison to the response time to the real-time probes. Second, the participants were asked to complete the NASA–TLX workload survey at the end of each scenario.

3. Performance measures: Appropriate performance-based measures of SA were collected for the surveillance (time to validate a track, accuracy in track assignment, and time with track lost) and identification (time to classify a target) positions. Because the link between SA and performance is not as direct for the weapons team, no direct performance measures were assessed. In addition, a subjective measure of performance was collected from the SMEs at the end of the study through a process in which the SMEs rank ordered the teams on the basis of performance.

Procedures

Each team participated in two simulation trials. Half of the participants began with the peace scenario (average workload), and the other half began with the

TABLE 7
Study 2: Situation Awareness Global Assessment Technique Queries

Surveillance

1. On the attached map, indicate the targets that need symbology, special tracks, and unknowns.
2. Which noninitiated tracks, special tracks, or unknowns have changed heading?
3. Which noninitiated tracks, special tracks, or unknowns have changed code/mode?
4. Which tracks have emergencies?

Identification

1. On the attached map, indicate the pending targets.
2. Which targets need ID?
3. For the targets in No. 2, how much time is remaining for ID?
4. For the targets in No. 2, what is the code?
5. For the targets in No. 2, what is the direction of flight?
6. For the targets in No. 2, what is the speed?
7. For the targets in No. 2, what is the altitude?
8. For the targets in No. 2, what is the agency responsible for its airspace?

Weapons team

1. On the attached map, indicate the target aircraft's location.
2. For the aircraft in No. 1, what is its speed?
3. For the aircraft in No. 1, what is its heading?
4. For the aircraft in No. 1, what is its altitude?
5. For the aircraft in No. 1, what is its flight size?
6. Is the aircraft suspect by customs?
7. On the attached map, indicate the mission aircraft's location.
8. For the aircraft in No. 6, what is its speed?
9. For the aircraft in No. 6, what is its heading?
10. For the aircraft in No. 6, what is its altitude?
11. For the aircraft in No. 6, what is its type?
12. For the aircraft in No. 6, what is its frequency?
13. For the aircraft in No. 6, what is its callsign?
14. Is weather an impact on its route?
15. Are voice communications okay on its route?
16. How much time is available on fuel remaining?
17. Are there any hazards or emergencies that will affect performance?
18. Is the aircraft within 5 miles of its airspace boundaries?
19. Distance to inner ADIZ?
20. What is the range and bearing from the mission aircraft to the target aircraft?
21. Where is the intercept point?
22. How much time to intercept?
23. Is an intercept over water possible?
24. What is the mission type?

Note. ID = identification; ADIZ = Air Defense Identification Zone.

war scenario (high workload). Before each trial, participants were provided with instructions regarding the simulation, SAGAT, real-time probes, the secondary task, SART, and the NASA-TLX rating form. Six SAGAT freezes, 16 real-time probes insertions, and 12 secondary task measures were administered during each scenario. The SAGAT freezes and real-time probes were arranged so that they did not occur within 2 min of each other. At the end of each scenario, the participant completed the NASA-TLX and SART rating forms. In addition, at the end of each trial the SMEs completed an Observer-SART rating for the participants. At the end of the study, the SMEs performed the rank-ordering task.

Results

The data collected in this study allowed for two assessments of the real-time probe: the sensitivity of the measure to scenario differences and the validity of the measure to accurately reflect SA.

Sensitivity

To ascertain the level of sensitivity possessed by each metric, ANOVAs were performed between the war and peace scenarios for each measure.

Real-time probes. Two aspects of the real-time probe measure were analyzed for sensitivity to scenario differences: response accuracy and response time. The real-time probes were evaluated at two levels: the individual query level and at an overall level (created by combining the responses across each of the positions into one large category and then analyzing the category as a single item). When individual queries were examined within each position for response accuracy, only one showed sensitivity across scenarios: the weapons team “fighter location” probe, $F(1, 6) = 31.24, p = .001$. The Overall measure, on the other hand, showed significant sensitivity between scenarios, $F(1, 151) = 9.78, p = .002$.

When the individual queries were evaluated for sensitivity of response time to the correctly answered probes, only two of the real-time probes showed sensitivity to scenario differences: the identification “code” probe, $F(1, 5) = 6.91, p = .047$, and the weapons team “target heading” probe, $F(1, 11) = 4.41, p = .060$. The Overall measure for response time showed marginal sensitivity to scenario differences, $F(1, 226) = 3.72, p = .055$.

SAGAT. When the SAGAT data were evaluated, the ANOVAs indicated that 11 of the 21 queries showed significant sensitivity to scenario differences ($p < .05$): 2 of 4 surveillance queries, 2 of 2 identification queries, and 7 of 16 weapons team queries (see Table 8). The aggregated SAGAT data did not show any sensitiv-

TABLE 8
Results for Situation Awareness Global Assessment Technique Analyses of Variance

	<i>df</i>	<i>F</i>	<i>p</i>
Overall	(1, 1,832)	0.353	.553
Surveillance			
Emergency	(1, 55)	2.316	.134
Symbology*	(1, 54)	4.299	.043
Heading*	(1, 55)	4.555	.037
Code	—	—	—
Identification			
Time remaining*	(1, 57)	8.436	.005
Code*	(1, 57)	8.610	.005
Weapons team			
Target altitude*	(1, 100)	10.767	.001
Fighter altitude	(1, 101)	0.000	.996
Target speed*	(1, 100)	14.537	.000
Fighter speed	(1, 101)	2.347	.129
Callsign	(1, 102)	0.249	.619
Frequency	(1, 102)	0.226	.635
Fighter location*	(1, 72)	5.529	.021
Target location	(1, 63)	3.465	.067
Weather*	(1, 102)	8.572	.004
Time to intercept*	(1, 78)	9.653	.003
Target heading	(1, 100)	0.952	.332
Fighter heading	(1, 101)	3.297	.072
Playtime*	(1, 100)	8.139	.005
Intercept point	(1, 33)	0.451	.507
Flight size*	(1, 100)	4.678	.033
Fighter type	(1, 102)	0.122	.727

*Statistically significant.

ity, as was expected because SAGAT queries tend to be independent of one another. Individual SAGAT data provide more useful information and, in this case, showed sensitivity to scenario differences for each of the three positions.

It was interesting that not all of the queries showed higher SA in the peace than the war scenario; rather, trade-offs in SA were observed. For example, in the war scenario, the weapons team showed less SA with regard to target altitude, target speed, flight size, and playtime but showed greater SA with regard to fighter location, weather, and time to intercept. Such trade-offs in attention to different aspects of SA are common (Endsley, 1995b) and illustrate the diagnostic power of the SAGAT measure.

SART. The results from the ANOVAs performed on the SART data showed that the measure was significantly sensitive to scenario differences for the weapons

team position, $F(1, 18) = 5.63, p = .029$, and for the Overall measure (i.e., SART ratings combined across all four positions), $F(1, 38) = 4.31, p = .027$. The Observer–SART ratings were also assessed and showed sensitivity for the surveillance, $F(1, 5) = 11.85, p = .018$; weapons team, $F(1, 14) = 48.75, p < .000$; and overall Observer–SART ratings (i.e., Observer ratings combined across all four positions), $F(1, 29) = 11.8, p = .002$.

The means of the SART and Observer–SART ratings show SA to be rated lower in wartime scenarios than in peacetime scenarios. The only exception to this trend is for the identification position. Participant feedback for this position supported these findings; participants indicated that the war scenario posed no increase in difficulty level for them.

Workload. Analysis of the NASA–TLX workload ratings showed sensitivity to scenario differences for the weapons team, $F(1, 18) = 6.90, p = .017$, and for the Overall measure, $F(1, 38) = 7.88, p = .008$. This finding was similar to that found from the SART analysis. The analysis of the secondary task measure found no evidence that the measure was sensitive to scenario differences ($\alpha = .05$).

Performance Measures

Although several performance measures were originally planned, many of these encountered problems during data collection and had to be omitted from the analysis. Only two objective performance measures were available for analysis: total time to validate a track (surveillance) and mean time to classify a target (identification). These measures were analyzed via ANOVAs: Neither of these performance measures showed sensitivity between scenarios. These results were not surprising because, as with many domains, sensitive and meaningful performance measures are difficult to find.

Summary

The objective of the sensitivity analyses was to determine if the real-time probes possessed sensitivity comparable to other measures utilized during the study. Several of the measures demonstrated sensitivity to scenario differences (e.g., real-time probe, SAGAT, SART, NASA–TLX), thereby indicating that a true difference did exist between the scenarios. On this point, the results for the real-time probe are mixed. Although both real-time probe accuracy and response time showed sensitivity when analyzed as an aggregate measure across all three teams, an examination of individual probe type showed little sensitivity. However, the trends in the means frequently followed the trends in the SAGAT data, thus suggesting that the lack of sensitivity may be due primarily to the comparatively small amount of data collected for each real-time probe type. Even though the

real-time probes were inserted 16 times during the scenarios (as compared to only 6 for SAGAT), only one probe at a time could be assessed (whereas the SAGAT technique allows all the queries to be provided at each freeze). Therefore, each probe type could only be provided between one and four times per trial (depending on the number of SA requirements for the position).

Validity

The validity of the real-time probe as a metric of SA was examined in two ways in this study: by comparing the real-time probe responses directly to SAGAT (which represents a validated measure of SA) and by comparing the real-time probe response time to the response time for the secondary task (to determine if SA or workload is being assessed by the measure).

Real-time probes versus SAGAT. Linear regressions were performed comparing real-time probe responses (accuracy and response time) to the corresponding SAGAT queries. For the comparison between real-time probe response time and SAGAT, the regressions for the Overall measure, surveillance, and weapons team were significant (see Table 9). For the comparison between real-time probe response accuracy and SAGAT, all the regressions were significant (see Table 10). Thus, a weak but significant correlation exists between both real-time probe measures and SAGAT, indicating that at least at some level, the real-time probes are indeed measuring SA.

Real-time probes versus secondary task. In accordance with the theory behind real-time probes, the responses time to the real-time probe may be indicative of the participant's SA in that a shorter response time would mean the participant was aware of the element, whereas a longer response time would indicate that the participant had to search the display for the answer. However, an alternate explanation for variations in response time would be that the response time to the real-time probes was a measure of spare capacity and were thus essentially a secondary measure of workload. To assess this possibility, the relationship between

TABLE 9
Regression Results: Real-Time Probe Response Time Versus Situation Awareness Global Assessment Technique

	<i>df</i>	<i>F</i>	<i>p</i>	<i>R</i> ²
Overall	(1, 221)	14.63	.001	.062
Surveillance	(1, 30)	5.21	.030	.148
Identification	(1, 22)	1.28	.270	.055
Weapons team	(1, 165)	11.83	.001	.067

TABLE 10
Regression Results: Real-Time Probe Accuracy Versus Situation Awareness Global
Assessment Technique

	<i>df</i>	<i>F</i>	<i>p</i>	<i>R</i> ²
Overall	(1, 258)	26.65	.001	.094
Surveillance	(1, 30)	5.77	.023	.161
Identification	(1, 41)	7.13	.011	.148
Weapons team	(1, 183)	11.53	.001	.059

the real-time probes and a secondary task was assessed. The mean secondary task response time, the mean real-time probe response time, and the mean real-time probe accuracy were calculated for each 10-min time block within each trial. Neither the linear regressions comparing real-time probe response time to the secondary task nor the linear regressions comparing real-time probe accuracy to the secondary task were significant ($\alpha = .05$). Thus, no evidence exists to suggest that the real-time probe is measuring simple reaction time.

Comparison of Measures

In addition to the regressions, a Pearson's correlation matrix was calculated to directly compare the various measures employed in this study. Several correlations were found, each in the expected direction.

- NASA-TLX was negatively correlated with the Observer-SART ratings ($R^2 = .45$, $p = .044$); higher workload was associated with a lower Observer-SART rating.
- Response time to the real-time probes was negatively correlated with the Observer-SART ratings ($R^2 = .51$, $p = .023$); longer probe response time corresponded to lower observer SART ratings.
- Response time to the real-time probes was positively correlated with rank ($R^2 = .411$, $p = .043$); longer response times to the probes were associated with worse overall rankings.
- Response time to the real-time probes was marginally correlated with the NASA-TLX score ($R^2 = .409$, $p = .073$); longer response time to the probes correlated with higher workload ratings.
- Response time to the real-time probes was marginally correlated with real-time probe accuracy ($R^2 = .41$, $p = .073$); longer response time to the probes was associated with increased accuracy.
- Real-time probe response accuracy was marginally correlated with NASA-TLX ($R^2 = .41$, $p = .072$); increased accuracy was associated with higher workload ratings.

Summary

The linear regressions suggest that real-time probes are weakly correlated with SAGAT and that real-time probes were not correlated with the secondary workload task. This evidence suggests that the real-time probes are indeed measuring at least some aspect of SA and are not simply acting as a secondary workload task. However, the Pearson's correlation matrix did indicate that the real-time probe was weakly correlated with NASA-TLX score, thereby preventing the exclusion of workload as a consideration in the attempt to validate real-time probes.

Summary Study 2

Unlike the previous study, both response time and accuracy of the real-time probes provided useful metrics. Both were sensitive to differences between the war and peace scenarios at the combined level but not at the individual probe level. This result indicates that for real-time probes to effectively differentiate between conditions, a larger number of repetitions is needed than occurred at the individual probe level. SART showed results similar to the real-time probes. Conversely, individual SAGAT queries showed sensitivity but the combined measure did not. This finding is not surprising, as SAGAT queries tend to be independent. A weak but significant correlation existed between the real-time probes and the corresponding SAGAT queries, suggesting that real-time probes were at some level providing a measure of SA. No correlation was found between the real-time probe and the secondary task workload, indicating that the real-time probe does not measure simple reaction time that is indicative of workload. However, a weak correlation was found with NASA-TLX, which prevents ruling out the possibility that the real-time probes were providing a measure of workload. Thus, the possibility that the secondary task measure used in this study did not adequately capture workload fluctuations (as evidenced by its lack of sensitivity to scenario differences) must be considered. Therefore, the relationship between real-time probes and workload warrants further examination. Finally, subjective reports reveal that the participants did not consider the real-time probes intrusive or bothersome, thereby suggesting their applicability to operational settings. The protocol utilized in this study (i.e., not administering the probe while the participant was verbally communicating) probably minimized the disturbance.

CONCLUSIONS

Although the first study found no evidence that the real-time probes measured SA, the second study found a weak correlation between real-time probes and the

corresponding SAGAT queries, thereby suggesting that real-time probes were indeed measuring some facet of SA. Thus, real-time probes warrant more investigation before being dismissed from consideration as a metric of SA. Along this vein, a few recommendations for further testing of the concept can be gleaned from the results of these studies. First, as evidenced by the sensitivity and validity of the real-time probes at an aggregate level but less so at the individual probe level in the second study, more probes need to be administered for the measure to attain a sufficient level of reliability. In contrast to SAGAT, which allows an entire battery of questions to be asked at each stop thereby providing numerous repeated measures for each query, the repeated measures for the real-time probe was limited by the nature of the probe; probes are asked one at a time at various points in the scenario. In each of the preceding studies, the real-time probes were administered in the same scenario as several other metrics. If only the real-time probes were administered, considerably more probes could be administered during a single trial. In addition, minimizing the number of real-time probes by restricting them to the most crucial information would also allow for an increase in the repeated measure.

Next, both studies found weak correlations between the real-time probes and at least one of the workload measures employed by the study—ATWIT in the first study and NASA-TLX in the second. However, study two provided for the direct comparison of reaction time to a secondary task with both reaction time and accuracy of the real-time probe but found no correlation, indicating that the real-time probes were not measuring simple reaction time. Thus, no clear answer was found regarding the exact nature of the relationship between real-time probes and workload, signifying that this issue needs further research.

The need for a measure capable of assessing SA during real-time activities is apparent. The preceding studies assessed the ability of the real-time probe to meet this need. Although real-time probes are far from being validated as a measure of SA, their potential is promising. Although real-time probes will not provide as much data as SAGAT, in situations where no simulation facility exists the real-time probes may provide a viable option for measuring the SA provided by a given design concept. Additional testing and analysis of the technique, based on the recommended modifications, should be conducted to assess further the utility of real-time probes.

ACKNOWLEDGMENTS

This work was supported in part by the Federal Aviation Administration Technical Center, under the direction of Earl Stein and Randy Sollenberger; and TRW Inc. and the Air Force Operational Test and Evaluation Center, under the direction of Jim Lozito and Al Diehl. We thank these individuals for their support in this work.

REFERENCES

- Durso, F. T., Hackworth, C. A., Truit, T. R., Crutchfield, J., Nikolic, D., & Manning, C. A. (1998). Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly*, 6, 1–20.
- Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd Annual Meeting* (pp. 97–101). Santa Monica, CA: Human Factors Society.
- Endsley, M. R. (1990a). Predictive utility of an objective measure of situation awareness. In *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 41–45). Santa Monica, CA: Human Factors and Ergonomics Society.
- Endsley, M. R. (1990b). *Situation awareness in dynamic human decision making: Theory and measurement*. Unpublished doctoral dissertation, University of Southern California, Los Angeles.
- Endsley, M. R. (1995a). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37, 32–64.
- Endsley, M. R. (1995b). Measurement of situation awareness in dynamic systems. *Human Factors*, 37, 65–84.
- Endsley, M. R., & Bolstad, C. A. (1994). Individual differences in pilot situation awareness. *International Journal of Aviation Psychology*, 4, 241–264.
- Endsley, M. R., Sollenberger, R., Nakata, A., & Stein, E. (2000). *Situation awareness in air traffic control: Enhanced displays for advanced operations* (Tech. Rep. No. DOT/FAA/CT–TN00/01). Atlantic City, NJ: Federal Aviation Administration William J. Hughes Technical Center.
- Gutman, J. A., Stein, E., & Gromelski, S. (1995). *The influence of generic airspace on air traffic controller performance* (Tech. Rep. No. DOT/FAA/CT–TN95/38). Atlantic City, NJ: Federal Aviation Administration William J. Hughes Technical Center.
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA–TLX (Task Load Index): Results of empirical and theoretical research. In N. Meshkati (Ed.), *Human mental workload* (pp. 139–183). Amsterdam: North-Holland.
- Sollenberger, R. L., Stein, E. S., & Gromelski, S. (1997). *The development and evaluation of a behaviorally based rating form for assessing air traffic controller performance* (Tech. Rep. No. DOT/FAA/CT–TN96/16). Atlantic City, NJ: Federal Aviation Administration William J. Hughes Technical Center.
- Stein, E. S. (1985). *Air traffic controller workload: An examination of a workload probe* (Tech. Rep. No. DOT/FAA/CT–TN84/24). Atlantic City, NJ: Federal Aviation Administration William J. Hughes Technical Center.
- Taylor, R. M. (1990). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In *Situational awareness in aerospace operations (AGARD–CP–478)* (pp. 3/1–3/17). Neuilly Sur Seine, France: NATOAGARD.

Manuscript first received June 2002