Assessing the Situation Awareness of Pilots Engaged in Self Spacing

Thomas Z. Strybel, Kim-Phuong L. Vu, Jerome Kraft & Katsumi Minakata

Center for the Study of Advanced Aeronautics Technologies
California State University, Long Beach
Long Beach, CA 90840

We measured situation awareness (SA) of pilots in a simulation of an approach to a large metropolitan airport (DFW), using both SAGAT and SPAM probe techniques. Both methods of SA measurement significantly predicted pilot performance on a self-spacing task but in SPAM scenarios, probe latency predicted IAS variability, and in SAGAT scenarios, accuracy predicted IAS variability.

INTRODUCTION

Under the Next Generation Airspace Transportation System (NGATS), airspace operators will assume new roles and responsibilities in reaction to several essential changes in air traffic management operations, and the introduction of new automation technologies to support these important developments. For these changes to succeed, it is important that the information required of each new operator role be identified and the impact of these changes on operator situation awareness (SA) and workload be assessed. Because existing measures of SA and workload have not been evaluated in this environment, either existing techniques must be re-validated, or new measurement techniques must be developed that are valid, reliable and sensitive to the changes in operator SA and workload caused by NGATS.

The construct of SA affects human performance in many complex systems. Intuitively, SA refers to one's comprehension of the situation, or simply knowing what's going on around you. SA affects human performance on tasks that have unpredictable inputs and require that responses be made either within a critical time period, or in proper sequence (i.e., before another response is made). Pilot and air traffic control (ATC) tasks are highly time-dependent, and it is not surprising that a great deal of research has been performed on SA of these operators. Intuitively, airspace operators know what is meant by SA: controllers refer to it as "having the picture;" and pilots have called it "staying ahead of the aircraft" (Endsley et al., 1998; European Air Traffic Management Programme, 2003).

Despite considerable research on SA, there is no universally accepted definition or model of the concept. The most widely cited definition of SA is Endsley's (see e.g., Endsley, 1995): SA is based on three element levels: the perception of information in the environment within a volume of time and space, the integration and understanding of the meaning of the information, and the projection of the information to future events (including knowing what further information is needed to maintain SA). In a recent Eurocontrol review, SA was defined as the continuous extraction of environmental information, the integration of this information with previous knowledge to form a coherent mental picture, and the use of that picture in directing further perception and anticipating future events (Jeannot, et al., 2003; Dominguez et al., 1994). Both definitions have been criticized, however, as not adequately reflecting the operator's view of SA. ATCs, for example, view SA as a prioritized list of elements in the airspace environment. Ruitenberg (1997) states that terms such as "elements" or "information" do not specify exactly the nature of the elements or information. Controllers believe that SA, in addition to traffic information, includes personal factors, weather, equipage, navigational aids and performance.

With these diverse definitions of SA, it is not surprising that standardized methods of SA measurement are unavailable. In the last two decades, many SA measurement methods have been advanced. In recent reviews of SA, Eurocontrol (2000) reviewed nine methods; Salmon et al. (2006) evaluated 17 methods. Nevertheless, it has been difficult to develop a measurement method that meets all of the psychometric and operational criteria of a good measurement tool (e.g., Salmon et al., 2006). Most SA measures can be classified into one of three categories, probe techniques, rating scales, or performance-correlated measures.

Probe techniques administer SA-related queries during a simulation. The most commonly used probe techniques are Endsley's "freeze-probe" technique, known as SAGAT, in which questions are asked during simulation pauses, and Durso's Situation Present Awareness Method (SPAM) in which individual questions are asked during the course of a simulation run without scenario pauses (e.g., Endsley, 1995; Durso et al., 1997). In Endsley's often cited studies of air-traffic-management SA (e.g., Endsley, Bolte, & Jones, 2003**,** SAGAT probes are typically presented during scenario freezes at random intervals. ATC is asked about operationally relevant aspects of locations and characteristics of aircraft in a sector. Pilots are typically queried about nearby aircraft. These questions are developed through Endsley's Goal Based Task Analysis Procedure.

With SPAM (Durso et al., 1997) operators are asked individual questions in the course of a scenario while performing normal tasks. According to Durso, with good SA, either task-relevant SA information is held directly in memory or the location of this information is held in memory. Therefore, in SPAM, SA is measured as both the number of correct responses and the time to answer the question correctly. If the information being queried is held in the operator's memory, he or she can respond quickly. If the information is not in memory, but available on a display, response time will be faster if the operator knows where to find the information. To separate the effects of workload from SA, the operator is usually asked if he/she is ready for a question, and the question is not asked until the operator responds affirmatively to the ready prompt. The time interval between the ready prompt and the operator's acceptance is taken as a measure of workload. Of course, presenting questions online during a scenario could interfere with task performance (possibly changing SA), so questions must be presented in a manner to be consistent with the task. For example, ATC probes can be asked via the controller's landline; pilot probes can be asked by a confederate copilot.

The questions utilized in SPAM also differ from SAGAT. Instead of questions on absolute traffic information (e.g., "What is the altitude of AAL45?"), SPAM queries are relative (e.g., "Which aircraft is higher?"). This is considered to be more compatible with how ATCs represent traffic information. Moreover, Durso et al (2006) suggests that knowledge of past information, in addition to present and future information is relevant to SA. SPAM has not been tested as thoroughly as SAGAT, so validity and reliability information are unavailable. However, SPAM reaction times predicted novice ATC performance after variance due to cognitive skills was removed. And, SPAM reaction times were correlated with Remaining Action Counts, a measure of efficiency in a simulated ATC task (Durso et al., 2004, 2006).

Although there remain differences between specific techniques, probes in general are promising methods for measuring SA. They have been shown to be sensitive to operator and task environment differences, and diagnostic information can be obtained regarding the cause of poor SA. However, most questions are chosen to be relevant to the task; in fact, Endsley recommends a Goal-Directed Task Analysis be performed in order to identify critical questions, yet this technique can be time consuming. Moreover, the focus on operational relevance limits their generalizability. Measuring SA as either the percentage of correct responses or reaction time to scenario-specific questions limits the usefulness of the technique to qualitative comparisons within an individual experiment instead of quantitative comparisons between different scenarios, operator roles, and dissimilar NGATS ATM concepts.

In the present investigation, we measured SA for pilots in a simulation of an approach to a large metropolitan airport (DFW), using both SAGAT and SPAM probe administration techniques. However, for both techniques we developed probe-question *categories* from the existing literature and subject matter experts, in order to empirically determine which combination of question categories are related to (and can subsequently predict) SA-related performance measures. Once a set of probe question categories is known to be related to performance, specific probe questions in each scenario could be tailored to the scenario, situation, automation concept, etc., based on the previously described information requirements analysis.

In a previous experiment (Strybel, et al., 2007) we showed that online probe questions were related to subjective SA as measured by a standardized SA rating instrument (Situation Awareness Rating Technique, or SART). Probe questions were administered during a simulation predicted SART scores administered at the end of a scenario. Surprisingly, we showed that accuracy of responses to probe questions was not related to subjective SA. Instead, pilot estimates of distance to a patrol vehicle in the vicinity, and ratings of threat of encroachment were significantly related to SART SA scores. In the present experiment, we compared questions regarding the pilot's task that were either delivered during a scenario freeze as in SAGAT, or delivered online during a scenario run, as in SPAM. In addition, we developed question categories for SAGAT and SPAM administrations that were based on previous assumptions of each technique and from our previous research. Two categories of questions were used, processing and time frame. Processing refers to the cognitive operations required to answer a question: recall, comprehension or subjective assessment. Time-frame refers to the time focus of the question: past, present or future.

METHOD

Participants.

Thirteen licensed pilots (all males) were tested. All pilot participants were VFR rated and indicated being at least somewhat experienced with IFR. On a seven point rating scale, with 1 indicating no IFR experience and 7 highly experienced with IFR, only one pilot rated his instrument flight experience as lower than 4 (somewhat experienced). The mean experience rating was 5.2 (SD=1.4). Seven pilots reported having experience with glass cockpits, with the number of glass cockpit hours ranging from 5 to 4200. The mean number of flight hours for all participants was 1648 (SD = 2356). Participation was voluntary, and participants were paid $20 per hour for their time.

Scenario.

The simulated environment employed in this experiment was the northwest arrival corridor to the Dallas Fort-Worth (DFW) airport, similar to that used in Strybel et al. (2007). All scenarios were developed in consultation with an Air Transport Pilot and former Southern California TRACON Controller. Pilots entered the scenario near to the top of descent, and ended near the top of ILS approach to runway 18R. Pilots flew either a modified BOWIE-9 or GREGS-5 STAR arrival, with area traffic merging from the BOWIE 9, GREGS 5 and MASTY 2 arrivals. These STAR arrivals represent the northwest approach corridor to the18R and 13R runways at DFW. Participants were always assigned the 18R runway as their destination; however, all scenarios ended near the top of ILS approach.

Within the first two minutes of each scenario, ATC assigned limited self-spacing responsibility to the pilot: ATC instructed the pilot to achieve 10 miles lateral separation from an assigned lead aircraft when ownship was 12 nm outside of the GIBBI fix (approx 17 minutes from beginning of scenario). Under these rules, the pilot was required to modify his/her own speed to ensure that the necessary spacing from an assigned lead plane at an assigned waypoint was achieved. Permissible speeds ranged from 170kts to the filed FMS speed for the active leg of the flight. In each scenario, the lead aircraft was in front of ownship and on ownship's route. Eight scenarios were developed, and these varied in difficulty. For easy scenarios, illustrated in Figure 1, the lead (+1) aircraft is slowing down in response to vectoring traffic, but remains on ownship's route throughout. For hard scenarios, illustrated in Figure 2, the lead aircraft merges into ownship's route and engages in multiple vectors throughout the scenario. During a scenario run, participants were responsible for following ATC issued commands, meeting self-spacing requirements, and obeying the rules of the road at all times.

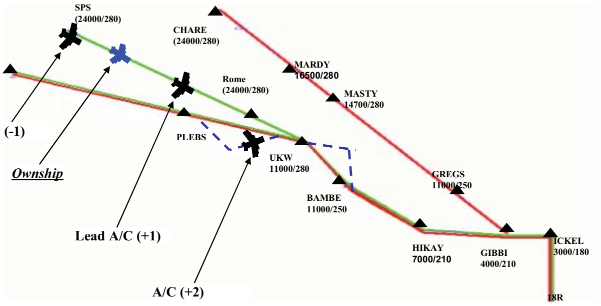

Figure 1. Easy Scenario Flight Path. The pilot's ownship follows the assigned lead (+1) aircraft. Lead aircraft never deviates from the arrival path.

During a scenario run, pilots were instructed to maintain self-separation from lead when directed by ATC, monitor and respond to ATC communications, identify and maintain awareness of lead AC, and monitor flight deck systems. ATC party line scripts, developed by a former SoCal TRACON Controller, were read by student ATCs and pilots and included directions to the participant.

Apparatus.

Simulations were conducted in CSAAT at California State University, Long Beach (CSULB). The simulation software, Multi-Aircraft Control System (MACS), Aeronautical Datalink and Radar Simulator (ADRS), Cockpit Display of Traffic Information (CDTI) and DAGVoice were developed by NASA Ames AOL and FDDRL laboratories (Prevot et al, 2005; Canton, 2006). Pilot participants used the MACS software in single pilot mode, and had the following flight instruments available: mode control panel, flight management system, primary flight display, and, in SPAM runs, a datalink tool. Participants used the CDTI in 2-D mode as a substitute for Traffic Collision Avoidance System (TCAS) display. ADRS acted as a communications hub and radar emulator for the simulation. DAGVoice served as a voice-IP-based party line tool for realistic ATC communication and chatter.

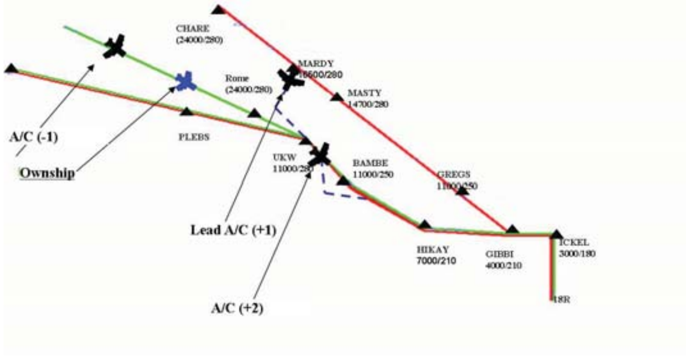

Figure 2. Hard Scenario Flight Path. Pilot's ownship follows an assigned lead (+1) aircraft that merges into ownship's route and engages in multiple vectors during the scenario.

For scenarios using SPAM queries, a DataLink tool was provided. Ready Prompts and Questions appeared in the DataLink window, accompanied by an audio alert. SPAM queries were sent to the pilot from a MACS station configured as TRACON Controller. Precise timing and accuracy of question delivery was facilitated by preprogramming the questions on a programmable keypad. In effect, a complete Datalink message could be sent with a single key press. For SAGAT queries, the scenario was paused, displays turned off, and a paper-pencil questionnaire was administered.

Design.

The experiment employed a 2 (scenario difficulty: easy, lead aircraft not vectored vs. hard, lead aircraft vectored) X 2 (probe administration method: SAGAT vs. SPAM) within-subjects design, creating four treatment combinations: Hard scenario (vectoring lead) -SAGAT, Easy Scenario (non-vectoring lead) - SAGAT, Hard scenario (vectoring lead) - SPAM, and Easy scenario (non-vectoring lead) - SPAM. Participants flew in two similar scenarios for each condition, for a total of 8 scenarios, presented in random order.

Procedure.

Each participant completed one four-hour training session, and one four-hour testing session. Training consisted of a two-hour briefing on the airspace, task and SA measurement methods, followed by two hours of practice on similar scenarios. In test sessions, each participant flew eight scenarios. SA probe question batteries (SAGAT-administrations) or a series of individual SA probe questions (SPAM-administrations) were administered throughout the simulation. SPAM queries were administered every two minutes beginning one minute into the scenario. These individual probe queries were administered via the datalink capability of the MACS software. A SAGAT test battery of eight questions was administered twice during each SAGAT scenario at roughly 6.5 and 13.5 minutes. SAGAT batteries were administered via pencil and paper tests, with the simulation paused and displays turned off. At the completion of each scenario run, pilots completed the NASA-TLX workload and SART 3D SA questionnaires. At the end of the experiment, pilots rated the information contained in the questions for relevance to the self-spacing task.

Pilot performance on the self spacing task was measured in terms of variability in Indicated Air Speed (IAS; Casso & Kopardekar, 2001). We assumed that pilots with high SA would make fewer and smaller adjustments in IAS. We also, recorded the number of missed pilot responses to ATC directions and the number of ATC directions that had to be repeated. SA was assessed by the percentage of correctly answered questions for SAGAT scenarios, and by the percentage of correct answers and response latencies (time between question presentation and pilot response) in SPAM runs. In addition, readiness response latencies (time between “ready question” prompt and pilot’s acceptance) were recorded as a measure of workload (Durso, 2006).

Question Development.

The SA questions used in SAGAT and SPAM scenarios were similar. They were chosen to determine if certain categories of information were more predictive of SA than others. The categories selected were developed from the SA literature and pilot ratings of information relevance. The question categories tested here were as follows.

- *Time Frame.* These questions queried the pilot on past, present or future events in the scenario.
- *Processing*: These questions were classified as recall, comprehension or subjective assessment. Recall questions asked for information held in memory or, in SPAM scenarios, found on cockpit displays. Answers to comprehension questions required that pilots process the information either held in memory or on cockpit displays. Subjective assessment questions required pilots to rate the threat of encroachment of nearby aircraft.

RESULTS

To determine if our scenario manipulations were effective and examine the performance effects of SA-administration condition, a two-factor repeated measures analysis of variance was run on the IAS standard deviations, with the factors scenario difficulty and SA method. Main effects of scenario difficulty, F(1,13) = 46.4; p<.001, and SA method F(1,13) = 6.20; p=.03, were obtained, as shown in Figure 3.

Greater IAS variability was observed for hard scenarios, suggesting that our difficulty manipulation was effective. Greater variability also was found for SPAM compared with SAGAT scenarios at both difficulty levels, suggesting that SPAM administrations did interfere with performance to some extent. Similar analyses were run on NASA TLX workload scores and SART SA ratings and these showed no significant effects of scenario difficulty or SA administration method.

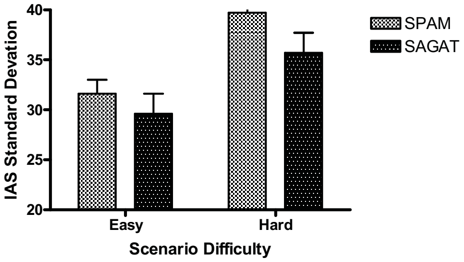

Figure 3. Standard deviations of IAS as a function of scenario difficulty and SA administration method.

SPAM

All correlations between measures of pilot experience and speed variability were nonsignificant. SPAM SA and workload latencies were log transformed to ensure normally distributed variables prior to analysis.

Table 1. Correlations between potential predictor variables and pilot performance in SPAM scenarios.

Predictor Variable	Standard Deviation IAS	Missed ATC Directions	Repeated ATC Directions
NASA TLX	.197	-.007	.239
SART	.049	.083	-.115
SPAM Workload Latency	.089	.197	.221
SPAM SA Latency	.281 *	$.265^{?}$	.119
SPAM Percent Correct	-.035	.02	.005

* p < .05 $^{?}$p=.07

Table 1 shows the correlations between potential predictor variables (SPAM SA latencies, SPAM workload latencies, NASA TLX, and SART), and flight performance variables (speed variability, missed ATC directions and repeated ATC directions). SPAM SA latencies were significantly correlated with IAS variability, such that longer latencies were associated with higher variability. None of the remaining variables were correlated with IAS variability. The correlation between SPAM SA latencies and number missed ATC instructions was marginally significant. When these variables were regressed against IAS variability, only SPAM SA Latency was a significant predictor, (β= .36; p=.025; r^2 = .08). Significant correlations also were obtained between NASA TLX scores and SPAM Workload Latency (r(52) = .32; p=.02), suggesting that the time to accept the ready prompt was associated with workload.

We also examined whether certain categories of questions were more effective in predicting pilot performance. We examined accuracy and latencies for the overall categories (time frame and processing) without examining the interactions of these variables because of the small sample size. This, of course, creates substantial collinearity between categories, which limit our conclusions. Nevertheless, with respect to SA latency, the latency to subjective-assessment questions was significantly correlated with IAS variability (r(13)=.594, p=.03), meaning that long response latencies (indicating poor SA) were associated with greater IAS variability. Moreover, accuracy of comprehension questions was also significantly correlated with IAS variability (r(13)=-.64, p=.02). In this case, higher accuracy (meaning greater SA) was associated with lower IAS variability. Although these findings must be viewed with caution, the suggestion of a benefit for certain categories warrants additional investigation.

SAGAT

Table 2. Correlations between potential predictor variables and pilot performance in SAGAT scenarios.

Predictor Variable	SD IAS	Missed ATC Directions	Repeated ATC Directions
NASA TLX	-.201	.051	.19
SART	.227	.203	-.134
SAGAT Percent Correct	.268[?]	-.149	-.121

[?] p=.06

Table 2 shows correlations between predictor variables and pilot performance in SAGAT scenarios. The correlation between accuracy on SAGAT queries and IAS variability was marginally significant, r(52)=.268, p=.06. Accuracy on SAGAT questions was the only significant predictor of variability in IAS (β= .296; p=.03; r^2=.07). No significant correlations were observed between IAS variability were and SAGAT questions by category.

SPAM vs. SAGAT and Question Categories

When accuracy of responses was compared between administration conditions, interesting patterns emerged. Overall pilots were more accurate in SPAM scenarios (M= 42%) than SAGAT scenarios (M=31%). In SPAM administrations, greater accuracy was observed for present (M=51%) and future questions (M = 47%) than past questions (M=32%), probably because present and future (flight-plan) information is available on cockpit displays. In SAGAT administrations, however, accuracy for present questions (M= 38%) was higher than for past (M=29%) and future questions (M=29%), reflecting added burden on the pilot's memory. Similarly, in SPAM scenarios, more recall questions (M=66%) were correctly answered than comprehension questions (M=26%), but in SAGAT scenarios there was no difference between recall and comprehension questions (M = 33% and 31%, respectively).

DISCUSSION

Both SPAM and SAGAT measures of SA predicted pilot performance on a pilot self-spacing task, suggesting that either method may be suitable for assessing SA. For SPAM, only latency significantly predicted IAS variability. SPAM apparently interfered with task performance to some extent, as evidenced by the significantly higher IAS variability in these scenarios. For SAGAT scenarios, accuracy significantly predicted IAS variability. Accuracy may have been more effective here because more questions were administered. Also, although SAGAT measures were less intrusive from a performance standpoint, several pilots commented on the disruptions created by scenario pauses. In SPAM scenarios, comprehension questions were correlated with IAS variability, which is consistent with Durso's (2006) views on SA probes.

In SPAM scenarios, subjective assessment latencies were significantly correlated with IAS variability, similar to the findings of Strybel et al. (2007). Note, however, that both SAGAT and SPAM SA measures were not related to SART ratings of SA, contrary to Strybel et al. (2007). This may be due to differences in the length of the scenarios.

REFERENCES

Canton, R., Refai, M., Johnson, W. W., & Battiste, V. (2005). Development and Integration of Human-Centered Conflict Detection and Resolution Tools for Airborne Autonomous Operations. *Proc. 15th Intern. Symp. Aviat. Psych*. Oklahoma State University

Durso, F. T., Bleckley, M. K., & Dattel, A. R. (2006). Does SA add to the validity of cognitive tests? *Human Factors*, 48, 721-733.

Durso, F.T., Truitt, T.R., Hackworth, C.A., Crutchfield, J.M. & Manning, C.A. (1997). En route operational errors and situation awareness. The *International Journal of Aviation Psychology*, 8 (2), 177-194.

Endsley, M.R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors*, 37(1), 65-84.

Endsley, M. R., Bolte, B., & Jones, D. G. (2003). Designing for situation awareness. London, UK: Taylor and Francis.

Jeannot, E., Kelly, C., & Thompson, D. (2003). *The Development of Situation Awareness Measures in ATM Systems* (HRS/HSP-005-REP-01). Bretigny-sur-Orge, France: EUROCONTROL Experimental Centre.

Prevot, T., Smith, N., Palmer, E., Mercer, J., Lee, P., Homola, J. & Callantine, T. (2006). The Airspace Operations Laboratory (AOL) at NASA Ames Research Center. *AIAA Modeling and Simulation Technologies Conference and Exhibit* (AIAA 2006-6112). Keystone, Colorado.

Rodgers, M.D., Mogford, R.H., & Mogford, L.S., (1997). The relationship of sector characteristics to operational errors, *Air Traffic Control Quarterly*, Vol. 5, No. 4, 241-263.

Ruitenberg, B. (1997). Situational awareness in ATC: a model. The Controller, Vol 36, No.1.

Salmon P., Stanton, N., Walker, G., & Green, D. (2006). Situation awareness measurement: A review of applicability for C4i environments. *Applied Ergonomics* 37, 225–238

Strybel, T. Z., Vu, K.-P. L., Dwyer, J. P., Kraft, J., Ngo, T. K., Chambers, V., & Garcia, F. P. (2007). Predicting perceived situation awareness of low altitude aircraft in terminal airspace using probe questions. J. Jacko (Ed), *Human-Computer Interaction, Part I*, *Lecture Notes in Computer Science* 4550 (pp. 939–948). Berlin: Springer-Verlag.

ACKNOWLEDGEMENTS

This research was supported by NASA cooperative agreement NNA06CN30A. We thank Jack Dwyer of The Boeing Company, Ken Wells and Tom Morris for advice and feedback on scenarios and performance measures. We also thank Tom Prevot, NASA Ames AOL and Walter Johnson, NASA Ames FDDRL, for their donation of simulation software and continued technical support.