COMPARING SITUATION AWARENESS MEASUREMENT TECHNIQUES IN A LOW FIDELITY AIR TRAFFIC CONTROL SIMULATION

Russell S. Pierce*, Thomas Z. Strybel, and Kim-Phuong L. Vu****
***University of California, Riverside, **California State University, Long Beach**

Keywords: *Situation Awareness, Air Traffic Control, Simulation*

Abstract

Changes to the National Airspace System will occur as part of the Next Generation Air Transport System. The consequences of these changes on the situation awareness of pilots and air traffic controllers need to be identified. Before situation awareness metrics can be deployed towards this goal, metrics that can unambiguously capture baseline levels of situation awareness in the current day National Airspace System must be developed. Towards the ends of developing such a measure we investigated the usefulness of three existing measures of situation awareness, a subjective technique, an offline probe technique, and an online probe technique, in predicting performance on a low fidelity air traffic control simulation. Existing situation awareness metrics demonstrated a capacity to predict performance metrics relating to airspace safety, but none were able to effectively predict important elements of airspace efficiency, viz. enroute and handoff delay. In some cases, the subjective technique and online probe technique interacted with a measure of workload.

1 Introduction

Previous researchers have shown that situation awareness (SA) is related to the number and severity of air traffic controller operational errors. Thus it is considered a critical variable in determining air traffic controller performance [1] [2] [3]. SA has been identified as a critical factor in determining the success of new air traffic management concepts [4]. However, all existing SA methods have limitations that prevent accurate predictions of air traffic controller (ATC) performance. The difficulty in capturing SA arises in part because the concept of SA itself is poorly understood and defined, and its relationship to ATC performance is complex. The need for valid SA measure is important for evaluating future air traffic management environments such as that envisioned in the Next Generation Air Transportation System (NGATS).

Improvements in air transport system efficiency and safety are primary goals of NGATS. Because NGATS will most likely rely on automation to increase the traffic density, the roles and responsibilities of air traffic controllers may change. Currently, a controller's job primarily consists of the following global activities: Detect and resolve conflicts, monitor and direct traffic to maintain a consistent and maximum flow, respond to pilot requests, and provide additional services such as weather and traffic advisories. These tasks require active ATC involvement in managing air traffic, and the level of involvement may change if new automation concepts are introduced to reduce controller workload. For example, the automation could monitor traffic and detect conflicts. Once the conflict is detected, the automation could either alert the controller to resolve the conflict, suggest a resolution for the conflict, or even resolve the conflict automatically without controller input. Although the introduction of this automation seems simply to reduce a controller's task load (e.g., remove the monitoring traffic component), it is more likely that the automation will change the nature of the controller's job. For example, in order for the controller to resolve a conflict being highlighted by the automation, s/he may need to assess the traffic proximal and more distal to the point of loss of separation. This new task may demand greater awareness of more complex traffic patterns before generating potential solutions or accepting automated solutions. The controller may need to form a different picture of the situation compared to current traffic management pictures, causing unknown changes in ATC's perception, comprehension, and prediction of elements in their operational environment, i.e. Situation Awareness (SA, [1]).

Clearly, NGATS automation concepts need to be evaluated for its impact on SA prior to deployment. Unfortunately, there is not yet a globally accepted standard by which to measure SA, and each existing method has been shown to have specific advantages and disadvantages (see e.g., [5]). The goal of the present study was to empirically compare the predictive validity of three types of SA measurement methods for determining how SA contributes to performance on the Air Traffic Scenarios Test (ATST), a low fidelity ATC simulator. The present report is based on a portion of a larger investigation that examined the relationship between SA and other cognitive variables (for details regarding the role of cognitive variables in this experiment see [6]). Here, we present the results of an empirical test of three SA measurement techniques, the Situation Assessment Rating Technique (SART), the Situation Awareness Global Assessment Technique (SAGAT), and the Situation Present Assessment Method (SPAM).

2 Method

As part of a wider exploratory study of SA participants came to the simulation center on ten separate days. On Day 1 we collected demographic information, conducted cognitive testing, and oriented participants to the simulation task they would be doing, the Air Traffic Scenarios Test (ATST). On Days 2-4, participants were trained in the ATST in two 20 minute sessions, and then completed additional cognitive tests.

Days 5-9 were test days where we collected the information relevant to the current investigation. On each of these days participants completed a task in addition to performing on the ATST simulation. They would either engage in secondary task while operating the simulation or complete a post-simulation SA questionnaire. The secondary tasks included a SPAM-style online situation awareness assessment, a list memory task, and a word shadowing task. The post-task questionnaires included the SART and a SAGAT-style offline situation awareness assessment. We used a partial latin squares design to ensure that an equivalent number of subjects completed any given task on any given experimental day.

On a test day, upon arriving to the simulation center participants were told what their additional task would be. Participants then completed a 20-minute practice ATST trial with the additional task. After a five minute break, they completed a 20 minute ATST test trial and their additional task. All participants, regardless of condition also completed the NASA Task Load Index (NASA-TLX [7] [8]) at the end of each session.

On the final day, participants completed a final cognitive test, were thanked, and debriefed.

2.1 Participants

Twenty-one participants (*M* age = 23.81 years, *SD* = 6.05 years), 7 males and 14 females, completed all phases of the study and were paid $100 on completion. None of the participants had previous experience with ATC tasks.

2.2 Simulation and Performance

We used a low fidelity air traffic control simulator that was developed by the FAA for controller aptitude screening, the Air Traffic Scenarios Test (ATST). Performance on the ATST has been previously used with SA measures to evaluate the correlation between SA and performance [9]. In this simulation, participants guided icons representing aircraft as quickly and safely as possible from a starting position to a specific destination. Valid destinations included four sector gates at the cardinal points of the display, and two airports, one located in the top half and the other in the bottom half of the display. These aircraft traveled at one of three speeds (Fast, Medium, or Slow) and one of three altitudes (Sector Gate Exit Altitude, Mid-Level Altitude, or Landing Altitude).

Participants were instructed on "rules of the road," which explicitly informed them about how to move aircraft through the airspace and separation requirements (e.g., maintain 5-nm lateral separation and 1 altitude-level of vertical separation). In order to correctly exit, an aircraft needed to be traveling at a specific altitude, speed, and heading when they arrived at the destination. Participants controlled aircraft with a computer mouse. Clicking on an aircraft would select it to receive the next command. Commands were issued by clicking command buttons on the right side of the display. The commands consisted of changes in altitude, speed, and heading. Aircraft would immediately, and with 100% accuracy, respond to these commands. The display updated aircraft positions in 7 second intervals to simulate radar sweeps. Scenarios started with five aircraft already in motion, and new aircraft would appear in grey at the periphery of the simulation space at a steady rate of one per 30 seconds. A participant would accept a new aircraft into the scenario by clicking on its icon.

An examination of the FAA *Air Traffic Control Specialist Performance Measurement Database* [10] in conjunction with the simulation output revealed seven performance variables of interest that were measured for each participant and scenario. Three of these (ATC Procedural Errors, ATC Violations and Collisions) are related to system safety. The remaining measures (Handoff Delay, Enroute Delay, Number Commands Issued, and Number Correct Exits) assess system efficiency.

- *Number of ATC Procedural Errors.* The number of aircraft that either arrived at an incorrect destination or at the correct destination at an incorrect speed, altitude, or heading.
- *Number of ATC Violations.* The number of airspace violations, for example when aircraft lost separation with boundaries, airports, or other aircraft.
- *Number of Collisions.* The number of times that aircraft ran into boundaries, airports, or other aircraft.
- *Total Handoff Delay.* The total number of seconds that aircraft were left at the periphery awaiting handoff acceptance.
- *Total Enroute Delay.* The total number of seconds that aircraft were in flight minus the minimal amount of time it would have taken them to fly in a direct line from their starting position to their destination position at maximum speed.
- *Number of Commands Issued.* The total number of commands issued in a scenario. In the context of the ATST, commands only could be given to a selected aircraft. In order to terminate the selection of an aircraft it is necessary to give them a command, and aircraft are automatically deselected following a command.
- *Number of Correct Exits.* The total number of aircraft that arrived at the correct destination going at the proper speed, level, and heading.

2.3 Situation Awareness Measures

This experiment examined three situation awareness assessment techniques, SART, SAGAT, and SPAM.

2.3.1 Situation Awareness Rating Technique (SART)

The SART questionnaire requires participants to rate demand on attentional resources, supply of attentional resources and understanding of the situation on a 1-7 scale. Responses to the SART result in a subscale for each of the aforementioned dimensions as well as a combined score based on the difference between attentional demand and the sum of supply and understanding ratings [7]. However, because it is a subjective measure concerns have been expressed that it is overly related to workload [11].

2.3.2 Situation Awareness Global Assessment Technique (SAGAT)

In SAGAT, the simulation is paused and the display is blanked while questions regarding the situation are asked. These questions are usually derived from a goal-directed task analysis [1]. Once a participant answers the questions, the simulation is then resumed only to be stopped again at some later point for additional SAGAT probe questions. In the present design, fifteen SAGAT-like probes questions were presented at the end of the 20 minute scenario; consequently, neither a simulation pause nor resumption of task activities was required.

2.3.3 Situation Present Awareness Method (SPAM)

A SPAM query began with a verbal “Ready” prompt presented via headphones. The participant indicated acceptance of the “Ready” prompt by pressing a spacebar. A verbal SA question was administered immediately following this key press. The participant then verbally answered the question. The first prompt for readiness arrived at 3 minutes into the scenario. Subsequent prompts arrived every 2.83 minutes, for a total of six prompts per scenario. In SPAM two SA measures are collected, the total number of correct responses, and the average amount of time it takes to respond correctly to a probe question, known as SPAM latency. In order to reduce variability due to differing question lengths, SA response latencies were measured from the offset of the question to the onset of the response on correct answers only.

3 Results and Discussion

3.1 Calculations and Transformations

A natural log transformation was used to normalize response latencies, ATC procedural errors, ATC violations, and collisions for the sake of statistical computation, but all values are reported in untransformed terms for ease of interpretation. For all ANOVAs, where appropriate, we report the original degrees of freedom but use the Huynh-Feldt correction for significance levels to account for violations of sphericity. All regression analyses reported here were conducted in 3 steps. First, the NASA-TLX was added to the model to predict a performance variable. Second, a situation awareness measure was added to the model. Third, the interaction between the situation awareness measure and the NASA-TLX was added to the model. All models were inspected for outliers visually and using dfBeta criteria. We report only those models which, after any outliers are removed, improve on the previous step to at least a marginally significant extent ($p \leq .10$).

3.2 Practice Effects

Our study was designed to eliminate practice effects by having participants complete training trials prior to our test days. Repeated measures ANOVAs demonstrated no practice effects for any performance measure across time, *p*s < .25. Although performance did asymptote on the ATST after 5 days of training, on average participants still committed 7 separation violations per test scenario.

3.3 Task Performance

We employed paired sample t-tests to determine whether there were any systematic differences in performance depending on the situation awareness measure being used. Performance was generally equivalent between SART and SAGAT, *p*s > .16. However, performance in the SPAM scenarios was lower than that obtained in the either the SART or SAGAT scenarios in terms of handoff delay, number of commands issued, and correct exits, *p*s < .03. As shown in Table 1, participants in SPAM scenarios issued fewer commands and made fewer correct exits. Moreover, the total handoff delay in SPAM scenarios was significantly longer than in either SAGAT or SART scenarios. These findings suggest that SPAM interfered more with performance on the ATC task than SAGAT and SART. In the following sections, each SA method was analyzed for its ability to predict the performance variables listed in Table 1.

Table 1: Means (and *SE*s) for the Seven Performance Variances for the Three SA Measures

Performance Variable	SART	SAGAT	SPAM
Handoff	2556.67	2491.00	3545.33
Delay	(501.37)	(613.72)	(630.43)
Enroute	6646.00	6798.42	6656.19
Delay	(243.08)	(237.27)	(198.74)
Procedural	2.00	1.43	1.81
Errors	(0.45)	(0.41)	(0.50)
ATC	6.95	5.95	8.24
Violations	(1.69)	(2.03)	(2.77)
Collisions	1.43	0.71	1.67
	(0.53)	(0.23)	(0.61)
Commands	214.90	215.76	202.48
Issued	(6.10)	(6.64)	(7.00)
Correct	32.35	33.05	30.62
Exits	(1.03)	(0.87)	(1.04)

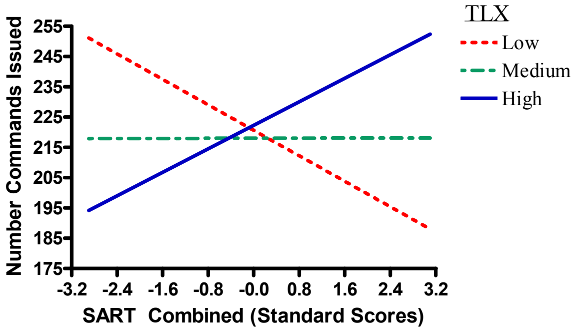
3.3 SART

SART was able to predict the number of procedural errors. Specifically, SART Combined and SART Understanding scores significantly predicted the number of procedural errors [$F(1, 17) = 4.58$, p = .05, $r^2 = .20$, and $F(1, 16) = 4.68$, $p = .05$, $r^2 = .22$, respectively]. The relationship between SART Supply and number of procedural errors was only marginally significant [$F(1, 16) = 3.37$, $p = .08$, $r^2 = .17$]. Each of these SART measures was inversely related to the number of procedural errors: high scores on SART were associated with fewer procedural errors. Additionally, higher SART Supply scores predicted that there would be fewer ATC violations, $F(1, 16) = 6.12$, $p = .02$, $r^2 = .24$.

SART Combined scores were not significantly related to number of commands issued, but the interaction of SART combined and NASA TLX workload was predictive, $F(1, 14) = 5.78$, $p = .03$, $r^2 = .30$, as shown in Figure 1. For individuals reporting low workload on the TLX, the number of commands issued was highest when SART-Combined scores were low, and number of commands decreased with SART Combined scores. In contrast, for individuals reporting high workload, increases in SART Combined scores were associated with more commands being issued. For individuals reporting medium levels of workload, the number of commands issued was constant across all SART SA ratings.

Of course, this interaction could also be interpreted in terms of the relationships between workload and number commands issued. When SART combined scores were high, workload was directly related to commands issued: high workload was associated with greater numbers of commands. With low SA combined scores, workload was inversely related to the number of commands issued. Given the subjective nature of both SART and TLX measures, it is difficult to separate the effects of perceived workload and SA on this performance measure. Since all participants encountered equivalent task environments, the variance in SART and TLX reflected mostly individual differences in evaluation of the task environment.

Fig. 1: Predictions of SART Combined Scores for Different Levels of Workload on Number of Commands Issued



It should be noted that after controlling for workload, the SART Demand scores did not predict any additional variance in performance (all *p*s > .19). This finding is not particularly surprising since the demand subscale is possibly most similar to the TLX workload scale.

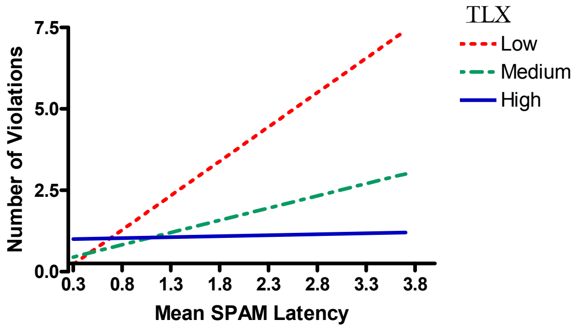
3.4 SAGAT

Overall accuracy on SAGAT probes was low (*M* = 5.19 or 35% correct; *SD* = 3.3 correct). Nevertheless, SAGAT was marginally related to the number of ATC procedural errors and had fewer violations, *F*(1, 18) = 4.09, *p* = .08, r^2 = .19 and *F*(1, 18) = 3.06, *p* = .10, r^2 = .12, respectively. Specifically, participants who scored higher on SAGAT committed fewer procedural errors and violations than those who scored lower on it.

3.5 SPAM

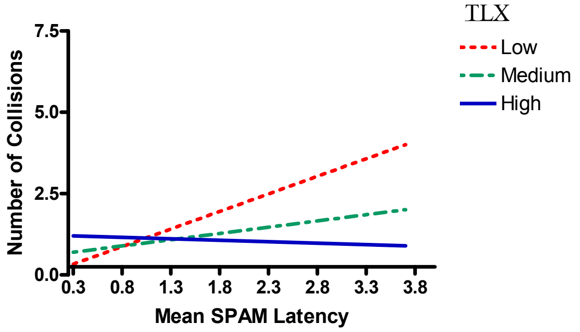
Overall accuracy on SPAM (*M* = 3.38 or 56% correct; *SD* = 1.20 correct) was relatively higher than on SAGAT, which is not surprising since the questions were administered individually during the course of the scenario and information was available on the ATC's display. Overall, participants took on average 2.68 seconds (*SD* = 1.84 s) to respond to SPAM probe questions. SPAM latency significantly predicted airspace violations, F(1, 18) = 10.86, p < .001, r^2 = .37. As shown in Figure 2, the interaction of SPAM latency with workload was significant, F(1, 17) = 8.22 p = .01, r^2 = .20. For individuals reporting low and moderate workload, increases in SPAM latencies were associated with increases in airspace violations. For participants reporting high workload, the number of violations was constant quite low. Similarly, with high SA, as determined by SPAM latencies, the number of violations was unaffected by workload. As SA decreased, the number of violations was determined by workload level.

Fig. 2: Predictions of SPAM Latency for Different Levels of Workload on Number of Airspace Violations



SPAM latency also predicted the number of collisions, *F*(1,18) = 4.39, p =. 05, r^2 =*.19.* Moreover, as shown in Figure 3, the interaction between SPAM latency and workload significantly predicted the number of collisions, F(1, 17) = 4.00, p = .06, r^2 = .16. This interaction followed the same pattern as number of ATC violations shown in Figure 2, which is expected from the relationship between ATC violations and number of collisions: collisions are ATC violations that were not resolved promptly.

Fig. 3: Predictions of SPAM Latency for Different Levels of Workload on Number of Collisions



Finally, regardless of reported workload, higher SPAM latencies were associated with fewer aircraft correctly departing the airspace, *F*(1, 14) = 6.72, *p* = .02, r^2 = .26. This finding indicates that lower SA leads to less efficient performance.

3.6 Summary

We examined the predictive validity of three existing SA measurement techniques using seven performance metrics that assess air traffic safety and efficiency. All three SA techniques did predict performance measures of safety to some extent. In terms of procedural errors, both SART and SAGAT were equally predictive, but SPAM latency was not predictive. For airspace violations, SART, SAGAT, and SPAM latency were predictive. Although SPAM (r^2 = .37) accounted for more variance than SART Supply (r^2 = .24), and SAGAT (r^2 = .12), these differences were not significant. However, the interaction of SPAM and workload together with the simple effect of SPAM latency was significantly more predictive of violations (r^2 = .58) than SAGAT, Z = 1.926, p = .05. Moreover, only SPAM latency predicted collisions, and the interaction between SPAM latency and workload further added to the predictive validity of SPAM (r^2 = .35).

In terms of performance measures of efficiency the evidence was mixed at best. Although the interaction between SART Combined and TLX workload did predict the number of commands issued, it is difficult to separate the effects of workload and SA when these two subjective scales are administered together. Only SPAM latencies significantly predicted the number of correct exits and this effect was not modified by workload. Lastly, none of these SA measures were able to predict enroute and handoff delays, both of which are related to efficiency.

We suspect that the problems in predicting efficiency reflect less on the measures of SA than on the performance metrics themselves. Each of the performance measures represents a non-independent component of a dynamic process based on ATC performance. For example, when an aircraft enters the scenario, it accumulates handoff delay until the participant takes the aircraft. At this point, the aircraft begins to accumulate enroute delay. Therefore, if the participant takes all aircraft as soon as they appear, handoff delay will be low, but increases in sector traffic will impact enroute delay. On the other hand, holding aircraft until they can be managed produces the opposite result. Future investigations into SA must utilize advanced statistical techniques, such as Principal Components Analysis, which will allow investigators to determine the most important elements of performance and clarify the relationship between operator performance measures and measures of system efficiency. Unfortunately such techniques were not available in the present investigation due to our sample size.

More research is required before prescriptions may be made with a high degree of certainty; however, on the basis of the available evidence, we suggest that SA in simulation environments be measured with a combination of SART and SPAM and TLX. Our experiment demonstrated that SPAM latency is not predictive when individuals report high workload. We suspect this is because participants neglected the SPAM probes in order to meet the demands imposed by the scenario. SPAM latencies in this condition may reflect both workload and SA, despite instructions to participants to delay accepting a question until time was available to answer it.

While generally desirable, this level of task involvement may be detrimental to the effectiveness of SPAM as a measure of SA. Consequently, scenarios, instructions, and probe rates should be adjusted based on anticipated workload levels. Since individuals may report varying degrees of workload, even given the same scenario and instructions, we also suggest that a measure of workload, such as the NASA-TLX, be administered at the end of the simulation task. And, experimenters may also wish to employ a combination of SA measurement techniques to assess convergent validity of the SA construct.

4. Acknowledgements

This research was supported by NASA cooperative agreement NNA06CN30A. We thank Jimmy Nguyen, Beau Krebsbach, Mary Ngo, Meghann Herron, and Katsumi Minakata for assistance with data collection and coding. For general advice and operational expertise we thank Jack Dwyer of The Boeing Company and

Tom Morris. For statistical advice we thank Chandra Reynolds, Timothy Gann, and the University of California, Riverside MAMA Statistics Forum. We also thank Dana Broach and FAA Civil Aerospace Medical Institute Training and Organizational Research Laboratory (AAM-520) for permission to use the ATST and technical support. Any opinions or conclusions are those of the author(s) alone, and may not reflect the official position or policy of NASA, Federal Aviation Administration or the U.S. Department of Transportation.

References

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

[2] Durso, F.T., Truitt, T.R., Hackworth, C.A., Crutchfield, J.M. & Manning, C.A. (1998). En route operational errors and situation awareness. *The International Journal of Aviation Psychology*, 8 (2), 177-194.

[3] 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, HCII 2007, Lecture Notes in Computer Science* 4550 (pp. 939–948). Berlin: Springer-Verlag.

[4] Next Generation Air Transportation System 2005 Progress Report. (2005). Joint Planning and Development Office.

[5] Jeannot, E. Situation Awareness, Synthesis of Literature Research. EEC 16/00. Eurocontrol Experimental Center.

[6] Pierce, R.S. Vu, K-P, and Strybel, T.Z. (in press). The Relationship Between SPAM, Workload, and Task Performance on a Simulated ATC Task. *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting*. New York, NY: Human Factors Society.

[7] Taylor, R.M. (1990). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. *Situational Awareness in Aerospace Operations, AGARD-CP-478*, 3-1 - 3-37.

[8] Hart, S.G. & Staveland, L.E. (1987). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. P.A. Handcock & N. Meshkati (Eds), *Human Mental Workload*. Amsterdam: North Holland Press.

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

[10] Hadley G.A., Guttman, J.A., & Stringer, P.G. (1999). Air Traffic Control Specialist Performance Measurement Database. DOT/FAA/CT-TN99/17

[11] Selcon, S.J., Taylor, R.M., & Kortisas, E. (1991). Workload or situation awareness?: TLX vs SART for aerospace systems design evaluation. *Proceedings of the Human Factors Society 35th Annual Meeting* (pp. 62-66). Santa Monica, CA: Human Factors Society

Copyright Statement