Optimizing Online Situation Awareness Probes in Air Traffic Management Tasks

Thomas Z. Strybel, Katsumi Minakata, Jimmy Nguyen, Russell Pierce, and Kim-Phuong L. Vu

California State University Long Beach,
Center for the Study of Advanced Aeronautics Technologies
1250 N Bellflower Blvd. Long Beach, CA 90840, USA
tstrybel@csulb.edu, kminakata@gmail.com, mrjimnguyen@gmail.com, russell.s.pierce@gmail.com, kvu8@csulb.edu

Abstract. We examined the effectiveness of situation awareness probe questions in predicting sector performance and behavior in a human-in-the-loop simulation air traffic management (ATM) simulation with low (50%) and high (75%) traffic densities. Probes were presented online during the performance of the air traffic management task, and the accuracy and response latencies were measured. Hierarchical linear modeling was used to analyze the predictive power of each category type. Response latencies for conflict probe questions predicted performance metrics associated with separation assurance.

Keywords: situation awareness measurement, air traffic management, NextGen.

1 Introduction

The most impacted operators in Next Generation Airspace Transportation System (NextGen) will be those of pilots and air traffic controllers (ATCs). Pilots operating in NextGen environments may assume expanded responsibility for flight planning and separation. ATCs will be using tools that enable them to safely and effectively share responsibility for separation assurance with aircrews and automation, while at the same time being centrally involved in managing aspects of new air-traffic-management (ATM) concepts. Presently, the impacts of these NextGen ATM concepts and technologies are unknown, yet success in meeting NextGen objectives depend on optimized function allocations between pilots, ATCs, and automated tools. Effective function allocation requires measures of operator situation awareness (SA), workload, and performance that can assess the impact of changing task demands. Unfortunately, reliable, valid, and robust measures are presently unavailable [1].

SA can be defined as either the processes used to develop and maintain awareness [2], or the information that determine the state of awareness [3]. A precise definition for the construct is still being debated, and consequently, there are no robust measures of the construct. SA measures usually fall into one of three types: subjective, performance-based, and probe. Probe measures query the operator about awareness of information. Two probe methods are commonly used: Situation Awareness Global

M.J. Smith and G. Salvendy (Eds.): Human Interface, Part II, HCII 2009, LNCS 5618, pp. 845–854, 2009.

Assessment Technique (SAGAT) [2], and Situation Present Assessment Method (SPAM) [3]. With SAGAT, an "offline" probe technique, the simulation is frozen, the operator's displays are blanked, and the operator is queried about information in the simulation. SPAM is an online technique in which probe questions are administered to operators individually during a scenario [3]. Durso et al. [3] showed that SPAM reaction times predicted novice ATC performance after variance due to individual differences in cognitive skills was removed. SPAM reaction times have been shown to be related to measures of ATC [3, 4] and pilot [5] performance. However, some investigations have found that online probing reduces performance and increases workload [4, 5].

One limitation of online probes is that a standard method for developing probe questions is nonexistent. For offline probes, a Goal Based Task Analysis Technique is recommended, but this technique is time consuming and focuses on information requirements without assessing either priorities or understanding of the task. Online probe questions are usually developed with subject matter experts, but information is needed on what (i.e., information content) to query and how (i.e., question format) to query in order for the technique to be useful in comparing NextGen concepts. In our previous investigations, probe questions addressed SA process (recall and comprehension) and time frame (past, present and future), but the content of information probed was not systematically manipulated [4, 5]. The present study examined the relative effectiveness of questions, based on types of processing, time frame, and information content, for predicting ATC performance variables. These categories were investigated in an ATM simulation in which ATCs managed traffic while responding to online probe questions.

2 Method

2.1 Participants

Seven students enrolled in the Aviation Sciences Program at Mount San Antonio College and nine retired air traffic controllers (6 TRACON and 3 ARTCC) participated in the simulation. For more information regarding participant background, see Vu et al. [6]. Each participant ran in six test scenarios with the order of scenario presentation counterbalanced between participants.

2.2 Apparatus

The simulation was run using the Multi Aircraft Control System (MACS) developed in the Airspace Operations Lab at NASA Ames Research Center. MACS is a medium fidelity simulation for simulating both ground and air side operations [7]. Each participant's ATC station was a simulated DSR display of combined sectors ZID 91 and 81. Simulated datalink and conflict probe tools were unavailable for ATC-pilot communications and conflict probing, although a simulated datalink window located outside of the DSR screen was used for online probing. Participant ATCs communicated with pseudopilots located in an adjacent room via VoiceIP software [8]. Six 40-minute scenarios were created, three of which approximated current-day low (50%) and high (75%) traffic densities. An automated ghost controller station

managed all traffic outside the participants' sectors, and initiated handoffs to the participant ATCs 15nm outside the sector boundaries. ATC participants, when appropriate, initiated AC handoffs to the ghost controller, which were automatically accepted after 30 seconds.

2.3 Procedure

Twelve probe questions were developed for each scenario. These were administered at three-minute intervals beginning at four minutes into the scenario. Probes were presented to participants in a datalink window located on the right side of the DSR display at roughly eye level. Participant responses were made with a CH Products Multifunction Panel that allows keys to be arranged in any order. Each key was programmed with a macro consisting of key presses, mouse movements and clicks to send a coded message from the probe display. Probe queries were administered by an experimenter located in an adjacent room. A probe sequence began with a "Ready Question" message sent to the participant's datalink window accompanied by an audio alert. When the participant had sufficient time to take a question, he/she pressed the Ready button sending an affirmative message back to the experimenter. The experimenter immediately sent the probe question, and the participant responded by selecting one of the six buttons located on the bottom of the response panel. If the Ready response was not acknowledged after two minutes, the query was withdrawn and the next probe was sent one minute later.

Queries were developed with subject matter experts who were familiar with the scenarios. The individual questions fit into one of three information processing categories, search/recall, comprehension and subjective assessment, and two time frames, immediate-past/present and future. Examples of questions fitting each combination of processing and time frame are presented in Table 1. Search/recall questions (e.g., Questions 1 and 2 in Table 1) could be answered by retrieving information from memory or finding information on the ATC display. No other processing was required to respond correctly. Comprehension probes (Questions 3 and 4 in Table 1) were used to assess the operator's understanding of the situation. Correct answers to these queries required the operator to retrieve information from memory or the display and process it. Subjective rating questions (Questions 5 and 6 in Table 1) were questions in which the participant provided an assessment of either the likelihood of an event or severity of a conflict. For each processing category, the probe question was directed at either the immediate past or present state of events, or required projection into the future.

In addition to the processing/time frame categorization, the content of probe questions addressed three areas of ATC task knowledge: Sector Status, Commands and Communications, or Conflicts (see Table 1). Questions on sector status requested information regarding current sector state, such as number of aircraft, number departures or distance to a boundary. Command/Communication questions probed ATCs knowledge of the next likely command to be issued, the last command issued, handoffs, and communication errors. Conflict questions probed knowledge of current and future conflicts between an aircraft pair. In addition to information contained in the question, the format of the questions was categorized as Multiple Choice, Yes-NO or rating. Multiple choice questions were answered by selecting one of six alternatives,

Table 1. Examples of probe queries and their classification based on processing (RC: Recall, CMP: Comprehension, SB: Subjective Assessment), time frame (IP: Immediate Past/Present, F: Future), Information Content (SEC: Sector Status, COM: Command/ Communication, CNF: Conflict), and question format (MC: Multiple Choice, OT: other).

	Processing & Time Frame						Information Content & Question Format					
	REC		CMP		SB		SEC		COM		CNF	
Sample Question	IP	F	IP	F	IP	F	MC	OT	MC	OT	MC	OT
1. How many AC are in descent to SDF NOW?	✓						✓					
2. Will FDX32 be the next overflight to exit your sector?		✓						✓				
3. How many pilot read back errors in the last 5 min.?			✓						✓			
4. How many conflicts will ASQ381 have if you take no further action?				✓							✓	
5. Rate concern about SWA2898 and AWE989.					✓							✓
6. Rate likelihood you will vector EGF494 for traffic.						✓				✓		

usually representing a quantity. For example a query "How many aircraft ..." was responded by selecting one of six response buttons labeled 0 thru 4, and 5+. Yes-no questions required answer of agreement/disagreement, and ratings were made on a six item scale corresponding to the six response buttons with the left-most button labeled "very low/very unlikely" and the right-most button "very high/very likely." Thirty-seven probes were multiple choice format, 17 yes-no and 18 rating. For subsequent data analysis, yes-no and rating questions were combined into an "Other" category. Unfortunately, the number of questions addressing each content area, processing category/time frame, and format combination was not equivalent. Therefore, each category was analyzed separately, and the interpretation of our results is limited to the effects of each probe category.

Participants' responses to probe questions were time stamped and saved in MACS data files. The correct answers for each scenario and participant were obtained by reviewing scenario video and audio recordings and MACs data files. The mean percent correct for probes were determined and averaged based on processing categories, time frame, information content and question format. Response latencies for correct and incorrect answers were also determined as a function of each category. These were analyzed as a function of participant group and traffic density. The results of participant experience are reported elsewhere in this volume [6]. The following ATC performance variables were analyzed:

- *Mean Handoff Time*: The average time per aircraft between accepting a handoff and handing it to the next sector.
- *Handoff Time Standard Deviation*: Standard deviation of handoff times for each participant and sector.
- *Mean Sector Time*: The average travel time through the sector per AC.
- *Sector Time Standard Deviation*: Standard deviation of sector times in a scenario.
- *Number LOS*. Total number of LOS per scenario.
- *Average Vertical Distance*: The average vertical distance between each aircraft pair.

From voice transcripts we obtained measures participant behaviors:

- *Percentage of altitude, heading and speed changes*: Relative number of changes made to aircraft in terms of altitude, heading and speed.
- *Number of Traffic Advisories*: Number of messages that pointed out nearby traffic.
- *Number of Corrections*: Number of times a corrections to an instruction was issued.
- *Total Number Communications*: Number voice messages sent by the ATC participant.

We examined the effects of probe categories on accuracy and latency, and the effectiveness of each probe measure in predicting ATM performance behaviors.

3 Results

3.1 Probe Performance

The percentage of correct responses and response latencies were analyzed with separate mixed ANOVAs with factors of experience, traffic density, processing, and time frame. A significant interaction of time frame and processing category was obtained for accuracy, $F(2,28) = 10.91$, $p < .001$. For the immediate time frame, participants showed most agreement with a subject matter expert in their assessment of the information being queried (see Table 2). For the future time-frame, accuracy was higher for recall and subjective assessment probes than comprehension probes. Significant effects of processing category, time frame, and their interaction, $F(2,28) = 21.08$, $p < .001$, were obtained on response latency. Latencies lowest for comprehension probes and immediate probes, but the latencies were more equivalent across processing categories for the future time frame. There is little evidence for speed-accuracy tradeoffs here, because for past/present probes the mean recall latency (13 s) was higher than the mean comprehension latency (9.1 s), yet accuracy was equivalent (56% vs. 58%).

Mixed ANOVAs evaluated probes based on information content and question format, see Table 2. A main effect of format, $F(1,14) = 49.1$, $p < .001$, and marginally significant interaction between format and information, $F(2,28) = 2.81$, $p = .07$, was obtained on the percent correct responses. As expected, the accuracy was significantly lower for multiple choice questions (M=55%) than to the other (yes/no or rating) questions (M=80%). Accuracy was similar among the information content categories for multiple choice questions, but highest for questions that probed sector status for the other format. There were significant effects of information content, $F(2,28) = 24.49$,

Table 2. Probe Accuracy and Latency For Each Probe Category

	Past/Present		Future			Mult. Choice		Other	
Processing	**PC**	**RT**	**PC**	**RT**	**Infor-mation**	**PC**	**RT**	**PC**	**RT**
Recall	56%	13 s	78%	16 s	Sector Status	51%	12 s	86%	16 s
Comp	58%	9 s	58%	12 s	Comm-and	56%	11 s	76%	12 s
Subj	89%	16 s	75%	15 s	Conflict	58%	14 s	78%	16 s

p <.001, format, $F(1,14) = 44.37$, $p < .001$, and a marginally significant interaction between them, $F(2,28) = 2.21$, $p = .08$, on response latencies. Response latencies for multiple choice questions were equivalent but latencies for the other format were 4 s faster for command probes.

3.2 Performance Measures

Table 3 compares the means of sector performance measures for low and high density scenarios. Although average handoff time per AC was not significant, the standard deviation of the handoff times was marginally significant. Greater variability for handoff times was shown in high density scenarios. The time through the sector was higher and more variable with higher traffic density. The average vertical distance between aircraft was higher in high density scenarios, but this difference only approached significance.

Table 3. Significant Effects of Traffic Density on ATC Performance Measures

	Low Density		High Density		
Performance Measure	**Mean**	**Standard Error**	**Mean**	**Standard Error**	**p**
Handoff Time Std Dev	145.0	8.0 s	170.2 s	9.7 s	.06
Sector Time	709.0	2.7 s	742.2 s	3.7 s	<.001
Sector Time Std Dev	191.0	4.8 s	195.0 s	3.3 s	<.001

Table 4 summarizes participant actions that were significantly affected by traffic density. The percentage of altitude and heading changes increased with density, while the percentage of speed changed decreased. Participants also issued significantly more traffic advisories in high density scenarios.

Table 4. ATC Behavioral Measures for Low and High Traffic Densities

	Low Density		High Density		
Measure	**Mean**	**Standard Error**	**Mean**	**Standard Error**	**p**
Altitude %	71%	3%	75%	2%	.08
Heading %	19%	3%	27%	2%	<.01
Speed %	6%	1%	2%	1%	<.01
Traffic	2.7	.4	4.3	.7	<.02

Some of these behavioral measures were significantly correlated with sector performance metrics. LOS was negatively correlated with number of traffic advisories, $r(89) = -.33$, $p < .001$, and positively correlated with number of corrections, $r(89) = .22$, $p = .04$. In effect, greater numbers of LOS were associated with fewer traffic advisories and more corrections. Variability in handoff times was positively correlated with percentage of heading changes, $r(89) = .29$, $p > .001$, and negatively correlated with percentage of altitude changes, $r(89) = -.24$, $p = .02$, and number of corrections, $r(89) = -.28$, $p < .001$. Greater variability in handoff times was therefore associated with more heading changes, fewer altitude changes, and fewer corrections.

3.3 Predicting Performance and Behavioral Measures from SPAM Probe Latencies

Hierarchical Linear Modeling (HLM) was used to determine the effectiveness of online probes in predicting performance and ATC behaviors. There are several advantages to this approach. HLM can be applied to unbalanced data, requires fewer assumptions about variance-covariance matrices, and with centered variables, HLM partitions variance into between-subject and within-subject components. Therefore, we evaluated the extent to which probe latencies predicted differences in performance between-participants and differences within-participants across the scenarios [9].

Probe latencies for each category were evaluated separately because of the unbalanced design. All response latencies were normalized by inverse transformations. For each predicted measure, an unconditional model having no predictors was developed. This model creates two intercepts, representing unexplained between-subject and within-subject variance. From these models, we determined that the relative proportion of between- and within-subject variance depended on the specific measure. For example, 71% of the total variance in handoff times was due to differences between subjects, but only 21% of the total variance in LOS was between subjects. After the Unconditional model was created, separate HLMs were created with two response latency predictors by centering predictor variables: a between subjects' predictor computed as differences in the mean probe latency of each participant from the grand mean, and a within subjects' component, computed as the differences in probe latencies between each probe latency and the participant's mean latency.

Table 5. Summary of HLM analysis of probe predictors

Measure	Intrasubject Probe Predictors	Intersubject Probe Predictors	Slope	p	Variance Reduction
Handoff Std Dev.	Future		.007	.04	6%
LOS	Conflict		-.04	.02	2%
	Multiple Choice		-.03	.02	2%
	Future		-.05	.01	2%
Ave Vert Distance	Conflict		-24.6	.03	9%
	Multiple Choice		-22.3	.05	15%
Altitude Change %	Command		-.004	.01	5%
Speed Change %		Subjective	.01	.12	1%
		Future	.008	.01	7%
Traffic Advisories	Conflict		.11	.05	10%

A summary of measures having significant predictors is shown in Table 5. Latencies of online future probes predicted handoff variability by reducing within-subject variance by 6%. Because of the inverse transformations, positive slopes means that the response latencies were inversely related to handoff time and standard deviation: Longer response times predicted lower standard deviations. Several probe categories significantly or marginally predicted LOS: sector status, conflict, multiple choice and future time frame, each reducing within subjects' variability 2%. For each probe category, the negative slope meant that faster response times were associated with fewer LOS. The average vertical distance was predicted by conflict, multiple-choice format and comprehension probe latencies, with faster response latencies predicting less distance between aircraft.

For behavioral measures of performance, the proportion of altitude and heading changes were significantly predicted by command probe latencies. Longer response latencies predicted a higher proportion of altitude changes and lower proportion of heading vectors. The proportion of speed changes was predicted by subjective probes and future-time-frame probes but these predictors reduced variance between participants. Traffic advisories were predicted by conflict-probes; longer latencies predicted fewer traffic advisories.

4 Discussion

This preliminary investigation of the efficacy of online situation awareness probes in predicting ATC behavior and performance suggests that the technique has merit and

may be used to predict SA and changes in SA when NextGen ATM concepts and automation tools are introduced. Online probe latencies for probes related to the ATC's awareness of conflicts significantly predicted the number of LOS; longer probe latencies for these questions were associated with greater numbers of LOS. The significant slope obtained with HLM was for intrasubject differences in probe latencies, suggesting that the probes are measuring changes within the operator over scenarios. Moreover, conflict probe latencies significantly predicted number of traffic advisories and number of corrections issued by ATC participants. This is not surprising when one considers that traffic advisories and corrections are negatively correlated with LOS. When ATC participants issued the most traffic advisories and the fewest number of corrections there were fewer LOS. Note also that sector status probes were significant predictors of LOS, possibly another component of SA is involved with LOS that is not determining number of traffic advisories. Similarly, command probe latencies were significant predictors of the percentage of altitude and heading changes. Faster probe latencies predicted a higher proportion of altitude changes and lower proportion of heading changes. Online probe latencies previously were shown to predict pilot error and novice ATC violations [4, 5].

Note that most significant predictor categories were based on information content. Very few significant predictors based on processing were found, and when these were significant, they were for between subject differences. For example, the proportion of speed changes was predicted by comprehension probes and future-time-frame probes. However, these slopes were for probe latencies averaged for each participant and centered on the grand mean. Possibly, these categories assess individual differences in ATC behavior, related to the cognitive skills identified by Durso et al. [3] as predicting performance. Caution must be taken, however, as the interdependence of categories makes definite statements difficult. Nevertheless, we believe these findings indicate that online probing as a method of measuring SA is promising.

Acknowledgements. This simulation was partially supported by NASA cooperative agreement NNA06CN30A.

References

1. Rantanen, E.: Development and Validation of Objective Performance and Workload Measures in Air Traffic Control. Tech. Report AHFS-04019/FAA-04-07. Univ. of Illinois, IL (2004)
2. Endsley, M.R.: Measurement of situation awareness in dynamic systems. Human Factors 37(1), 65–84 (1995)
3. Durso, F.T., Bleckley, M.K., Dattel, A.R.: Does situation awareness add to the validity of cognitive tests? Human Factors, 721–733 (2006)
4. Pierce, R.S., Strybel, T.Z., Vu, K.-P.L.: Measuring situation awareness and its contribution to performance in air traffic control tasks. In: Proceedings of the 26th International Congress of the Aeronautical Sciences, Anchorage AK (2008)
5. Strybel, T.Z., Vu, K.-P.L., Kraft, J.: Assessing the Situation Awareness of Pilots Engaged in Self Spacing. In: Proceedings of the Annual Meeting of the Human Factors and Ergonomics Society, pp. 11–15. HFES, NY (2008)

6. Vu, K.-P.L., Minakata, K., Nguyen, J., Kraut, J., Raza, H., Battiste, V., Strybel, T.Z.: Situation Awareness and Performance of Student versus Experienced Air Traffic Controllers. In: Smith, M.J., Salvendy, G. (eds.) Human Interface, Part II, HCII 2009. LNCS, vol. 5618, pp. 865–874. Springer, Heidelberg (2009)
7. Prevot, T.: Exploring the many perspectives of distributed air traffic management: The multi aircraft control system MACS. In: International Conference on Human-Computer Interaction in Aeronautics, HCI-Aero 2002, October 23–25. MIT, Cambridge (2002)
8. Canton, R., Refai, M., Johnson, W.W., Battiste, V.: Development and Integration of Human-Centered Conflict Detection and Resolution Tools for Airborne Autonomous Operations. In: Proceedings of the 15th International Symposium on Aviation Psychology. Oklahoma State University, Columbus (2005)
9. Singer, J.D.: Fitting individual growth models using SAS Proc Mixed. In: Moskowitz, D.S., Hershberger, S.L. (eds.) Modeling Intraindividual Variability with Repeated Measures Data: Methods and Applications. Lawrence Erlbaum, Mahwah (2002)