The Relationship Between SPAM, Workload, and Task Performance on a Simulated ATC Task

Russell S. Pierce[1], Kim-Phuong L. Vu[2], Jimmy Nguyen[2] & Thomas Z. Strybel[2]

University of California, Riverside[1]
Riverside, CA

Center for the Study of Advanced Aeronautics Technologies
California State University, Long Beach[2]
Long Beach, CA

Abstract. This study examined the influence of workload and dual-task performance decrements associated with the SPAM technique. Participants performed the *Air Traffic Scenarios Test (ATST),* which is a low fidelity air traffic control simulator developed by the Federal Aviation Administration. Performance on the ATST was followed by post-run questionnaires (baseline conditions) or in conjunction with SPAM queries, word shadowing lists, or memory lists (secondary task conditions). SPAM was no different from our baseline conditions in terms of subjective workload. In comparison to other secondary tasks, SPAM tended to yield workload levels intermediate to List Memory (high cognitive load) and Word Shadowing (low cognitive load). SPAM was found to lower performance relative to baseline conditions in three of the seven observed performance measures. These findings suggest that the use of a "ready" prompt for probe question administration is not sufficient for reducing performance decrements associated with secondary tasks.

INTRODUCTION

The construct of situation awareness (SA) has a long history (Landry, in press), and is considered an important factor that affects human performance in many complex systems, particularly those with time constraints such as the National Airspace System (NAS). Operators, pilots, and air traffic controllers, in the NAS have intuitive knowledge of the meaning of SA (e.g., European Air Traffic Management Programme, 2003), and the loss of SA has been shown to negatively affect performance in air traffic management systems (e.g., Durso et al., 1997). Despite the intuitive appeal of SA, it is a construct that is difficult to isolate and measure. One reason for the difficulty in measuring SA is that there is not a universal definition for it. Endsley's (e.g., 2000) often-cited definition of SA as being "the perception of elements in their environment within a volume of space and time, the comprehension of their meaning, and the projection of their status in the near future" alludes to components of SA, but has not led to the development of robust SA metrics despite years of research. As a result, SA has become a variable defined by subtraction, that is, it is defined by what it is not, rather than being based on explicit theoretical principles. The search for good SA metrics continues because they are needed to evaluate operator performance using current and future technologies.

In general, human factors metrics should possess a number of qualities, for example, high validity, reliability and sensitivity. In addition to these qualities SA metrics should also possess diagnosticity, generalizablity, and be accepted by operators. Typically subjective techniques have been shown to be acceptable by operators, generalizable, and at least moderately reliable, valid, and sensitive. However they typically can not identify the operational causes of poor situation awareness. In contrast, probe techniques should be diagnostic of the causes of poor situation awareness. However, while they have been typically shown to be sensitive, the reliability and validity data on them are mixed and they often suffer from low operator acceptance and poor generalizablity. In addition to the qualities listed above we suggest that good measures of SA should assess only SA, and not include other constructs known to affect performance, such as individual differences in cognitive capacity and workload.

A notable study by Durso, Bleckley, and Dattel (2006) showed that an online SA assessment technique, the Situation Present Awareness Method (SPAM), could predict performance in a cognitively oriented air traffic management task (ATST) after controlling for variance attributable to individual differences in cognitive abilities. This provides initial evidence that SA can be separated from other constructs. With the SPAM technique, operators are queried about SA-relevant information while simultaneously performing the task. Operators are first given a "ready" prompt to indicate that a question is available for administration. Operators are instructed not to accept the ready prompt until they are available to answer the question. Using this technique, SPAM is supposed to be able to distinguish between the effects of workload and SA. The latency between the presentation of the ready prompt and the operator's acceptance should be related to workload. Since the question is administered as soon as the operator indicates "Ready", the time between the presentation of the question and the operator's response is an indicator of SA. That is, if the operator was aware of the information being queried, s/he would be able to answer the question in a shorter time than if s/he were unaware of the information and needed to search for it.

However, Durso et al. (2006) did not report a separate measure of workload in their study, so it is unclear whether the influence of workload was indeed separated from SA. Moreover, in air traffic control (ATC) situations, such as the one examined by Durso et al, the situation unfold dynamically and a participant who was ready in one moment may find they are not ready by the time the question is fully asked. In other words, during the time in which the question is administered, the participant may find that two aircraft are in conflict, or will be in conflict, and resolves the conflict prior to answering

the question. The longer response latency, in this case, should be partially attributed to workload rather than the operator not being aware of the information being requested. Thus, it is possible that workload was increased when SPAM queries were delivered, despite the fact that the operator signaled readiness. If so, in the context of Durso et al.'s experiment, SPAM may not have successfully isolated SA from workload. Further, with the SPAM technique, it is assumed that by asking questions only when the participant is ready, primary task performance will be unaffected. Although Durso et al. did indeed find no significant differences between task performance in their Control and SPAM conditions, there was a trend towards poorer performance in the SPAM condition in every variable they examined. Thus, it is possible that SPAM does induce some dual-task performance decrements.

The purpose of the present study was to directly examine the influence of workload and dual-task performance decrements associated with the SPAM technique. To this end we used the same low fidelity ATC simulator and a similar SPAM technique for assessing situation awareness as employed by Durso et al. (2006). In addition, we included two measures of workload (NASA-TLX and SPAM readiness latency) and two secondary-task measures of cognitive load.

If the assumptions underlying SPAM are correct, there should be no difference in workload and performance between SPAM and baseline conditions. If the assumptions underlying SPAM are not met, performance on SPAM trials should be more closely related to performance on trials in which other cognitively demanding secondary tasks were included.

METHOD

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.

Procedure

This experiment took place in the context of a wider exploratory study of SA (for other information related to this study, see Pierce, Strybel, and Vu, 2008). The full experiment took place over ten days. On the first day participants filled out a demographic questionnaire, completed a cognitive test, and received orientation and training on the Air Traffic Scenarios Test (ATST).

On days 2-4, participants practiced the ATST task in two 20 minute sessions, and then completed another cognitive test. On days 5-9, participants performed the ATST under five conditions: SPAM, one of two secondary tasks (Word Shadow or Word List) or with two post-task questionnaires. The exact order in which they received these tasks was determined using partial latin squares For each condition, participants completed one practice ATST session for 20 minutes, in order to acclimate to the task. After a five minute break they completed another 20-minute ATST test in one of the five conditions. A computer presented SPAM questions and secondary task items and recorded the verbal responses of the participants.

The results of the post-task questionnaires are not reported here, but ATST performance in these conditions was used as baseline (Baseline 1 and Baseline 2) measures of performance. All participants, regardless of condition, completed the NASA Task Load Index at the end of each session (NASA-TLX; Gawron, 2000; Hart & Staveland, 1987; Nygren, 1991). On the final day, participants completed a final cognitive task, were thanked, and debriefed.

Materials

Air Traffic Scenarios Test (ATST). The ATST is a low fidelity air traffic control simulator developed by the Federal Aviation Administration as a screening tool for ATC applicants. It has been previously used to investigate SA in college students (Durso et al., 2006). 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 top half and the other in the bottom half of the display. These aircraft travel at one of three speeds (Fast, Medium, or Slow) and one of three altitudes (Sector Gate Departure Altitude, Mid-Level Altitude, or Landing Altitude). In order to correctly reach the destination, aircraft needed to be traveling at a specific altitude, speed, and heading before exiting the gate or landing. 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 every 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* (Hadley, Guttman, & Stringer, 1999) in conjunction with the simulation output revealed seven performance variables of interest:

- *Handoff delay* was the number of seconds that aircraft were left at the periphery awaiting handoff.
- *Enroute delay* was the 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.
- *ATC Procedural Errors* occurred when aircraft either arrived at an incorrect destination or at a correct destination with incorrect speed, altitude, or heading.

- *ATC Violations* occurred when there were violations of airspace, for example when an aircraft nearly hit boundaries, airports, or other aircraft.
- *Crashes* occurred when aircraft ran into boundaries, airports, or other aircraft.
- *Commands Issued* occurred whenever an aircraft was selected because in ATST, commands can be issued only after selecting an aircraft. The aircraft is automatically deselected following a command.
- *Correct Exits* occurred if an aircraft arrived at the correct destination going at the proper speed, level, and heading.

Situation Present Awareness Method (SPAM). A SPAM query began with an audio "Ready" prompt presented via headphones. The participant indicated acceptance of the "Ready" prompt by pressing the spacebar. An audio SA question was administered immediately following this key press. The participant then verbally answered the question. For example questions asked included "How many planes are headed towards a sector gate at the wrong speed to exit" and "What is the level of the last plane you gave a command to". The first prompt for readiness arrived at 3 minutes into the scenario; subsequent prompts arrived every 2.83 minutes afterwards, for a total of six prompts per scenario. In order to reduce variability due to differing question lengths, SA response times were measured from the offset of the question to the onset of the response on correct answers only.

Word shadowing and list memory. The word shadowing and list memory tasks were presented to participants six times during a trial using the same method employed with the SPAM queries. The participant's goal in the word shadowing task was to repeat each word as soon as it was presented. Each query had eight words. The participant's goal in the list memory task was to listen to an entire list of four-nine words. At the end of the list presentation, a double tone sounded, signaling the participant to repeat the words they remembered in the same order as they were presented. The words in both tasks were matched in terms of length and usage frequency.

NASA task load index (NASA-TLX). The NASA-TLX is a six item scale where participants rate subjective workload on six dimensions. Each item corresponds to a particular element of workload and these values combined create a total workload score (Gawron, 2000; Nygren, 1991).

Readiness latency. In those conditions where online probes were given, SPAM, Word Shadowing, and List Memory, we measured the time from the onset of the "Ready" prompt until participants pressed the spacebar. This was intended to provide an additional measure of workload.

RESULTS

Calculations and Transformations

A natural log transformation was used to normalize response latencies, ATC Procedural Errors, ATC Violations, and Crashes. 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.

Effects of SPAM on Workload

Differences in workload between SPAM, Baseline 1, and Baseline 2 test conditions were evaluated with separate single factor within-subjects ANOVAs on the combined NASA-TLX scores and on each subscale. The effects of test condition were non-significant, all *p*s > .11. Differences in workload between SPAM, Word Shadowing, and List Memory Tasks were also evaluated with separate single factor within-subjects ANOVAs on the combined NASA-TLX scores and each subscale. Significant effects were observed for the overall NASA-TLX scores, $F(2, 40) = 6.69$, $p < .01$, partial $\omega^2 = .15$, Mental subscale, $F(2, 40) = 10.75$, $p < .001$, partial $\omega^2 = .24$, Temporal subscale, $F(2, 40) = 3.27$, $p = .05$, partial $\omega^2 = .07$, and Performance subscale, $F(2, 40) = 3.41$, $p < .05$, partial $\omega^2 = .07$. Follow up analyses were performed on those variables reaching significance with Bonferroni adjusted *p*-values (see Table 1 for group means and Table 2 for *p*-values and effect sizes).

To determine whether workload, as indicated by readiness latencies, differ between SPAM, Word Shadowing, and List Memory conditions, we conducted a single factor within-subjects ANOVA on these variables. Readiness latencies were not recorded for two subjects, so their data is not included in this analysis. Overall there was no significant difference between the SPAM ($M = 6.21$ s), Word Shadowing ($M = 5.92$ s), and List Memory ($M = 7.81$ s) conditions, $F(2, 36) = 1.18$, $p = .32$, partial $\omega^2 = .01$.

Table 1: Mean and SE (in parentheses) of TLX scores in SPAM, Word Shadowing and List Memory Conditions.

Measure	SPAM	Word Shadowing	List Memory
NASA-TLX	41.27	34.40	48.25
	(3.79)	(2.97)	(3.89)
Mental	10.57	8.95	13.67
	(1.14)	(1.11)	(1.09)
Temporal	9.24	7.05	9.67
	(1.28)	(1.13)	(1.29)
Performance	7.95	5.29	8.19
	(1.01)	(0.91)	(1.28)
Frustration	7.24	4.62	9.14
	(1.25)	(0.93)	(1.23)

Table 2: *p* values and effect size r^2 (in parentheses) of Bonferroni-adjusted comparisons between SPAM and Word Shadowing, SPAM and List Memory, and Word Shadowing and List Memory.

Measures Compared	SPAM & Word Shadowing	SPAM & List Memory	Word Shadowing and List Memory
NASA-TLX	.19	.33	.004
	(.16)	(.12)	(.41)
Mental	.31	.03	.001
	(.13)	(.28)	(.50)
Temporal	.05	1.00	.14
	(.25)	(.01)	(.19)
Performance	.14	1.00	.02
	(.18)	(.00)	(.30)
Frustration	.06	.20	.001
	(.54)	(.00)	(.52)

Effect of SPAM on Performance

The effect of test conditions on ATST performance measures were evaluated similarly. There were significant differences between conditions in terms of handoff delay, $F(2, 40) = 3.92$, $p = .03$, partial $\omega^2 = .08$, number of commands issued, $F(2, 40) = 5.06$, $p = .01$, partial $\omega^2 = .11$, and number of aircraft correctly exiting, $F(2, 40) = 7.05$, $p < .01$, partial $\omega^2 = .16$. No significant differences were observed in enroute delay, number ATC procedural errors, number of ATC violations, and number of crashes, Fs $(2, 40) \leq 2.58$, ps $> .10$. Follow up analyses were performed on those variables reaching significance with Bonferroni adjusted p-values (see Table 3 for group means and Table 4 for p-values and effect sizes).

Table 3: Mean and SE (in parentheses) of ATST performance measures in SPAM and Baseline Test Conditions.

Measure	SPAM	Baseline 1	Baseline 2
Handoff Delay	3545	2557	2491
(total seconds)	(630)	(501)	(613)
Number	202.20	214.90	216.50
Commands	(7.35)	(6.10)	(6.94)
Number of	30.60	32.35	33.05
Correct Exits	(1.09)	(1.03)	(0.91)

Table 4: p values and effect size r^2 (in parentheses) of Bonferroni-adjusted comparisons between SPAM and Baseline 1, SPAM and Baseline 2, and Baseline 1 and Baseline 2.

Measures Compared	SPAM & Baseline 1	SPAM & Baseline 2	Baseline 1 & Baseline 2
Handoff Delay	.09	.10	.99
(total seconds)	(.21)	(.21)	(.00)
Number	.05	.07	.99
Commands	(.26)	(.22)	(.01)
Number of	.05	.01	.88
Correct Exits	(.26)	(.41)	(.05)

To determine whether there were differences in performance between SPAM and secondary task conditions, separate one-way within subjects ANOVAs were performed on each measure of task performance. The only effect reaching statistical significance was number of aircraft correctly exiting, $F(2, 40) = 4.47$, $p < .02$, partial $\omega^2 = .10$. All other performance measures were not significant, Fs $(2, 40) \leq 2.30$, ps $> .12$. A follow up analysis of the number of aircraft correctly exiting, using a Bonferroni adjusted p-value, showed that significantly fewer aircraft exited correctly in SPAM ($M = 30.60$) compared with the List Memory condition ($M = 32.76$), $p = .05$, $r^2 = .05$. All other comparisons were not significant.

DISCUSSION

In this paper, we examined the relationship between SPAM, workload, and primary task performance. Specifically, we determined whether the addition of SPAM probe questions during the performance of a simple air traffic control task increased workload and whether SPAM probe questions interfered with primary task performance.

In terms of subjective workload measured with NASA-TLX, differences between SPAM and our baseline conditions were nonsignificant. However, significant differences between SPAM, Word Shadow, and List Memory conditions were observed on some subscales of NASA-TLX, as shown in Table 2. For example, ratings of Mental Demand in the SPAM condition were not significantly different from ratings in the Word Shadowing condition, and ratings in both conditions were significantly lower than ratings in the List Memory condition. This suggests that the increase in mental demand produced by SPAM may be minimal because the List Memory Task is high in cognitive load, and Word Shadowing is low in cognitive load. However, ratings of frustration levels in SPAM were not significantly different from ratings in the List Memory condition, suggesting that SPAM probes increased frustration to a level comparable with a high-cognitive-load task. It is difficult to determine the extent to which these differences are caused by the secondary tasks, however, since NASA-TLX, is administered at the end of a scenario and the effects of peaks in workload on overall ratings are unknown. Future researchers may wish to use an online measure of workload such as the ATWIT (Stein, 1985) at varying temporal distances from the query in order to capture the time course of workload in relation to query responding.

Nevertheless, differences in readiness latencies between test conditions were not significant, although trends in readiness latencies matched those of the NASA-TLX in showing that workload was highest in the List Memory condition, moderate in the SPAM condition, and lowest in the Word Shadowing condition. If readiness latencies do reflect workload, they may be less sensitive to changes in task load than NASA-TLX.

Finally, we obtained preliminary evidence that ATST performance in the SPAM condition was affected by the administration of SPAM probe questions. In three of the seven observed measures of performance (handoff delay, number of commands issued, and number of aircraft correctly exiting), performance was significantly poorer than in Baseline Conditions. Even in comparison to the List Memory task, which elicited higher workload ratings, significantly fewer aircraft exited correctly in SPAM. This suggests that the ways in which SPAM disrupts performance are more complex than a simple utilization of cognitive resources.

Taken together, our results indicate that SPAM's use of a "ready" prompt for probe question administration is not sufficient for reducing performance decrements associated with secondary tasks. Note, however, that our results were obtained with a sample of novice air traffic controllers in a simplified ATC task. The extent to which they generalize to experienced ATCs and realistic ATC scenarios is presently unknown.

ACKNOWLEDGEMENTS

This research was supported by NASA cooperative agreement NNA06CN30A. We thank Beau Krebsbach, Mary Ngo, Meghann Herron, and Katsumi Minakata for assistance in 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

Durso, F. T., Bleckley, M. K., & Dattel, A. R. (2006). Does situation awareness add to the validity of cognitive tests? *Human Factors*, Winter, 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. (2000). Direct measures of situation awareness: validity and use of SAGAT. In M.R. Endsley & D.J. Garland (Eds). Situation Awareness, Analysis and Measurement. New Jersey: Lawrence Erlbaum.

European Air Traffic Management Programme (2003). The Development of Situation Awareness Measures in ATM Systems. HRS/HSP-005-REP-01.

Gawron, V. J. (2000). *Human performance measures handbook.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hadley, G. A., Guttman, J. A., Stringer, P. G. (1999). Air traffic control specialist performance measurement database. DOT/FAA/CT-TN99/17. U.S. Department of Transportation, Federal Aviation Administration.

Hart, S. G., & Staveland, L. E. (1987). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock and N. Meshkati (Eds) *Human mental workload*. Amsterdam: Elsevier.

Landry, S. J. (in press). Human-computer interaction in aerospace. In J. Jacko & A. Sears (Eds.) *Handbook of Human-Computer Interaction, 2nd Ed*, Erlbaum: Mahwah, NJ.

Nygren, T. E. (1991). Psychometric properties of subjective workload techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors, 33,* 17-33.

Stein, E.S. (1985). *Air traffic controller workload: An examination of workload probe.* (Report No. DOT/FAA/CT-TN84/24). Atlantic City, NJ: Federal Aviation Administration Technical Center 2) http://acb220.tc.faa.gov/products/bibliographic/tn8424.htm

Pierce, R., Strybel, T. Z., & Vu, K.-P. L. (2008). *A Comparison of the Validity of Situation Awareness Measures in a Simulated Air Traffic Control Task.* Paper accepted for presentation at ICAS 2008.