Comparison of SAGAT and SPAM For Seeking Effective Way to Evaluate Situation Awareness and Workload During Air Traffic Control Task

Mitsuki Fujino, Jieun Lee, Toshiaki Hirano, Yuichi Saito, and Makoto Itoh
University of Tsukuba, Japan

Evaluation of air traffic controller's situation awareness (SA) is becoming important for managing air traffic control with the growth of air traffic. This study compared two SA evaluation methods: Situation Awareness Global Assessment Technique (SAGAT) and Situation Present Assessment Method (SPAM) to understand how these techniques affect controllers' predictability in different traffic density settings. Twenty students undertook simple air traffic control simulations by using both techniques. We investigated how these techniques affect their workload with Subjective Workload Assessment Technique (SWAT) and NASA-TLX. SWAT scores showed that high traffic density increased participants' workload, and extra workload was posed right after answering SA queries. NASA-TLX scores were larger when SAGAT was used than when SPAM was used throughout the simulation. We found that the workload with SAGAT interferes with main tasks more than that of SPAM. Query scores results suggested that SPAM is more predictive to the assessment of the controller's SA.

INTRODUCTION

Recently, the demand for air transport is increasing all over the world. In 20 years, the world's air passenger demand is expected to be approximately double that of today. This trend has raised a concern that current air traffic systems may not cope the future demand of air traffic (e.g., Ministry of Land, Infrastructure, Transport and Tourism of Japan, 2010). With respect to this problem, the investigation of air traffic managements has been widely explored in several aspects – e.g., training methods (Updegrove & Jafer, 2017). In particular, reducing human error has been a major concern for air traffic control (e.g., Davis et al., 1963). FAA (1990) reported that human error is the most critical factor influencing air traffic accidents. Additionally, Boeing (1985) described that errors caused by air traffic controllers is approximately 4.3% of accidents of all commercial aircrafts. When we consider the increased air traffic in future, the air traffic controller perhaps has difficulty in efficient air traffic managements even for traffic controllers who have an extensive training. Here, one substantial factor shown to influence their performance is situation awareness.

A myriad of human factors studies has investigated situation awareness (SA) in several domains, including aviation (Endsley, 1999), surface transportation (Merat et al., 2018), and nuclear power plant (Hogg et al., 1995). SA is defined as "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" (Endsley, 2000). For air traffic controllers, fast and accurate information processing is expected in a short period of time during dynamic and complex management tasks. The current task of air traffic controllers is to manage aircrafts as well as to predict the position of aircrafts for ensuring minimum safe spacing between multi-aircrafts. Based on the process, they have to give instructions to other operators. The high levels of situation awareness are important to comprehend the meaning and environment of situation and to carry out successful operations. That is, deteriorated or loss of SA negatively affect performance of the air traffic controller. Indeed, 88% of air traffic accidents are relevant to operators' SA (Endsley, 1995a).

Heavy workload is one of factors resulting decreased level of SA for operators. If a certain amount of workload is burdened, operators' SA will be degraded (Kaber & Endsley, 2004). The association between controllers' workload and SA has been widely treated in prior literatures (e.g., Athènes et al., 2002). For example, the finding by Edwards et al. (2017) described that workload is the most relevant factor affecting air traffic controllers' SA. As aforementioned, the current air traffic control task for flight safety and efficiency leads increased workload even for skilled controllers. Whereas examining whether or not they maintain high levels of SA during the task is a critical human factor problem, much of the research up to now has been descriptive in how to assess controllers' SA regarding not the future but the current air traffic condition.

There are various approaches for SA evaluation. For example, de Winter and his colleagues (2019) applied eye tracking for SA evaluation. Hall and Phelps (1983) asked operators to speak what they thought during tasks and judged their SA based on the answer. Additionally, Endsley (1988) investigated SA by asking participants the contents related to surrounding environments directly. However, the work of Rose et al. (2019) described that because SA evaluation by an utterance is dependent upon individual differences in levels of knowledge or skills, individual SA between operators could not be compared easily. Additionally, judging based on operators' behavior and performance of WHAT may be not facilitated accurately if the operators do not show performance as expected (Pritchett et al., 1995). With respect to such issues and future air traffic condition, direct and objective investigations by asking operators the contents which relevant to situations are considered as effective SA evaluation ways (Endsley & Garland, 2000).

Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1988) and Situation Present Assessment

Method (SPAM; Durso & Dattel, 2004) have been facilitated for objective SA evaluations (Endsley, 2019). Both SAGAT and SPAM requires controllers to answer questions about air traffic situations (past, present, and future) at a random time. SAGAT interrupts the operation at a random time, and the controller has to answer questions about traffic situations. SPAM uses the identical question to SAGAT and presents approximately one to three queries from those questions at random times while the operation is in progress. Previous studies using these methods in the domain of air traffic control have reported that using either SAGAT or SPAM does not affect the controllers' workload of control operations (Morgan et al., 2012). Further, SPAM negatively affects their workload and performance of control operations (Pierce, 2012). Debate continues about the effective techniques for SA evaluation (e.g., Kraemer & Süß, 2015).

In order to explore a better understanding of SA evaluation methods for the future air traffic situation, the current study attempted to address the debate about the two SA evaluation techniques: SAGAT and SPAM. We investigated influences of two methods on operators' SA and workload by using an air traffic control simulation. For further discussion of SA in different traffic conditions, air traffic density gradually increased with the course of simulation to assess following two hypotheses:

H1: Because SPAM provides queries with participants in real time, an extra workload is loaded when they answer the query. Therefore, SPAM may increase a greater workload than SAGAT.

H2: Because participants can search for answers after receiving queries in case of SPAM, so the correct answer rate of the query in SPAM may be higher than that of SAGAT.

METHOD

Participants

Twenty students (11 females) between ages of 19 and 25 participated in the current experiment (*M* = 21.2 years old, *SD* = 1.49 years). They received reimbursements for their participation. We obtained informed consents from all participants.

Apparatus

The current experiment was conducted by using a Fujitsu LIFEBOOK A577/RX laptop and a DELL S2318H LCD monitor (23 inch). The laptop simulated radars used in air traffic control (ATC). For accurate understanding of simulation, a high-resolution monitor was used. An ATC sector that participants should manage was inside of taped area in Figure 1. A voice recorder was facilitated to measure their scores of questionnaires in terms of situation awareness after a whole experiment.

Simulation Description

CARATS (Collaborative Actions for Renovation of Air Traffic Systems, Japan; Ministry of Land, Infrastructure, Transport and Tourism of Japan, 2010) Open Data 2016 were used in the current experiment. The data are public releases of part of air traffic control data by the Ministry of Land, Infrastructure, Transport and Tourism of Japan with the aim of developing air traffic systems and promoting research and development. The data include which the position of aircrafts and the time information at that time during 2016.

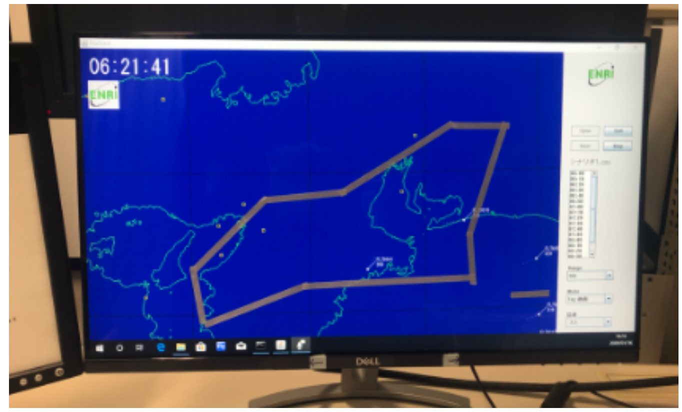

Figure 1. ATC simulation

PlotTrack_v2 software (Electronic Navigation Research Institute, Japan) can read this CSV file and display the aircraft with a white dot (Figure 2). In the current simulation (Figure 1), each white dot was updated every 10 sec, and the aircraft was displayed as if it were a video. In addition, the software provided 5 different playback speed options: speed x 5, x 10, x 20, x 50, x 100, and 5-times playback speed was used for the present study.

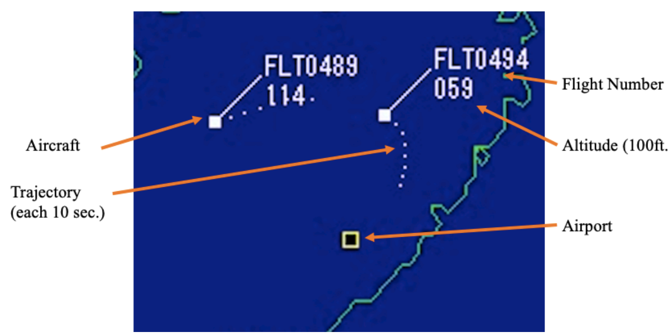

Figure 2. Information on the screen of PlotTrack_v2

Experimental design

A 2 x 6 repeated-measures experimental design with within-subject: SA evaluation method type (SPAM, SAGAT) and the measurement time of workload (S1–S6) and of queries (Q1–Q6) was used. All participants undertook the ATC simulation with both SA method types and were asked to answer levels of workload by using Subjective Workload Assessment Technique (SWAT; Reid et al., 1988) during the simulation.

In this simulation, the participant has to look at the aviation radar screen which displayed on the monitor and to keep tracking where aircrafts are located. If they predicted a dangerous approach between aircrafts on the way, they needed to note the time they noticed and the flight number of the relevant aircraft. We defined as a dangerous approach as when other aircraft approached within a radius of 5nm in the horizontal direction and within 2,000 ft in the vertical direction.

Because the experiment was conducted with not actual air traffic controllers but university students, complex tasks, such as making safe spacing, were not designed in the current study.

Dependent variables

Subjective Workload Assessment Technique (SWAT): In this study, simplified Japanese SWAT version (Miyake et al., 1993) was applied to measure the mental workload of the participants six times during the simulation. Participants weighted three items: Time Load, Mental Effort Load, Psychological Stress Load before an experiment. The final score was obtained during the simulation (ranged from 0 to 100).

NASA-TLX (Hart & Staveland, 1988): Simplified Japanese version NASA-TLX (Miyake et al., 1993) was used for assessment of participants' mental workload after the simulation of each condition. NASA-TLX consists of six items: Mental Demand, Physical Demand, Temporal Demand, Performance, Frustration, and Effort (ranged from 0 to 100).

Query score: SAGAT and SPAM evaluated the degrees of participants' SA at that time based on correct answer rate for each condition and query. In the current study, questions were designed by referring the work of Endsley (2019, p. 5) – i.e., "Mark all aircrafts on the indicated sector map". There are three questions for each query. A perfect score of a single query was 60 pt. (20 pt. for each question).

Procedure

Upon arrival, participants reviewed and completed a written informed consent at the beginning of the session. Participants were asked to weight three SWAT items. Then, the participants took two practices to be familiarized with the ATC simulation and each SA method type. Each practice lasted 10 minutes. After the practices, the participants completed two scenarios of ATC simulation in a randomized order. Each scenario lasted 15 minutes. Figure 3 illustrates time schedules of two scenarios for the current study. The scenarios were designed for reducing learning effect. In addition, we set three points for the examination (S2, S4, S5 in Figure 3; manipulated points), and the participants were asked to answer the SWAT immediately after answering SA questions. In both scenarios, air traffic density gradually increased.

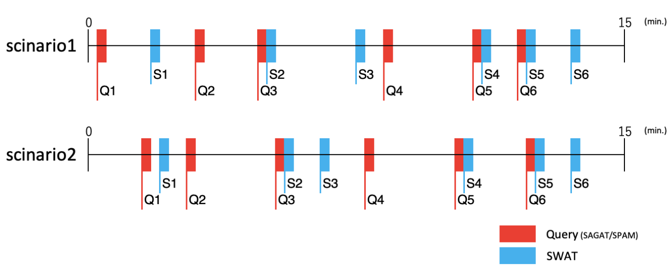

Figure 3. Scenario schedules

Statistical analyses

A 2 x 6 repeated measures Analysis of Variance (ANOVA) was performed to examine SWAT scores for each SWAT measurement time (S1–S6; see red lines in Figure 3) across SA evaluation method type (SPAM, SAGAT). The 2 x 6 ANOVA was also performed to analyse query scores for each query measurement time (Q1–Q6; see blue lines in Figure 3) across two method types. For post-hoc test, Bonferroni correction was utilized. The NASA-TLX score for each item as well as overall scores was analyzed by using a paired-samples t-test. All statistical analyses were implemented in IBM SPSS version 25.

RESULTS

SWAT

The mean SWAT scores obtained during the current experiment is shown in Figure 4. The 2 x 6 ANOVA observed a significant main effect on the SWAT measurement time [$F(5, 95) = 27.94$, $p < .001$, $\eta^2_p = 5.95$]. Table 2 describes the results of post-hoc test across the SWAT measurement time. The post-hoc test showed that the SWAT scores at 2nd, 4th, 5th, 6th points were significantly higher than those of 1st time point ($ps < .001$). Such tendency which SWAT scores increased with the experience of simulation was not observed from 4th point. Further, the SWAT scores at manipulated points were higher than other two points, 1st and 3rd (see Table 1 and Fig 4).

However, the effect of the SA evaluation method type and two factors did not interact ($ps > .645$).

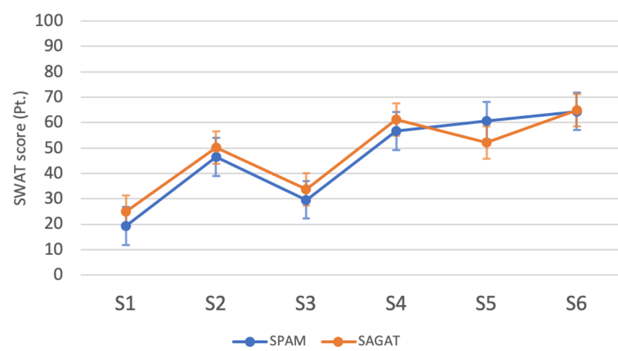

Figure 4. SWAT scores

Table 1. *p*-value of post-hoc test for SWAT scores
note. * means $p < .05$; ** means $p < .01$

	S1	S2	S3	S4	S5	S6
S1	-	**< .001	.818	**< .001	**< .001	**< .001
S2		-	** .004	.411	1.000	* .014
S3			-	**< .001	**< .001	**< .001
S4				-	1.000	1.000
S5					-	.597
S6						-

NASA-TLX

The paired t-test found a significant difference in NASA-TLX total score between SPAM and SAGAT [$t(19) = 2.70$, $p = .014$]. The total score of SPAM was significantly lower than that of SAGAT [M = 69.58 vs. 75.46]. Frustration of participants did differ across the question type [$t(19) = -2.58$, $p = .019$; M = 44.50 vs. 56.25]. The score of SPAM was significantly lower than that of SPAM for participants' frustration. The t-test observed favorable trends for the items of Performance [$t(19) = -1.76$, $p = .094$; M = 56.00 vs. 63.25] and Effort [$t(19) = -1.74$, $p = .099$; M = 76.75 vs. 82.00]. Consistently, the scores of SPAM was higher than SAGAT.

The remaining differences were not significant, $ps > .108$.

Query score

The total score of each query was 60 pt. (1 question = 20 pt.). A graph of the average query score obtained in an experiment is shown in the Figure 5. We performed ANOVA to analyse query scores with respect to two factors: SA evaluation method type and query measurement time. As a result, the main effect of the SA evaluation method type was observed [$F(1, 19) = 115.92$, $p < .001$, $\eta^2_p = .859$]. The query score of SPAM was significantly higher than that of SAGAT. The ANOVA also showed the main effect of the measurement time [$F(5, 95) = 27.95$, $p < .001$, $\eta^2_p = .595$]. Table 3 describes the results of post-hoc test for query scores across the measurement time. The query score at 6th time point was significantly lower than other points: 1st, 2nd, 3rd, 4th, and 5th point. However, there was no interaction between the two factors ($p > .202$)

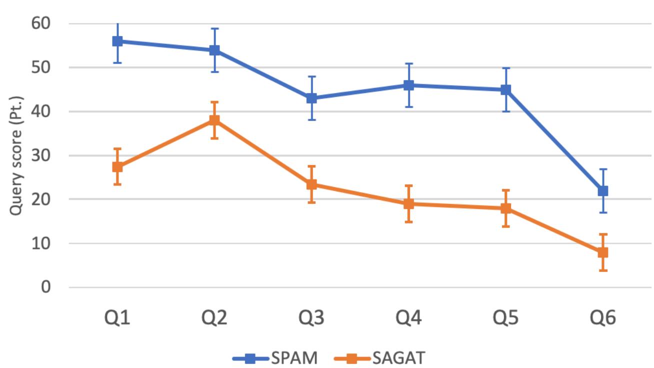

Figure 6. Query scores

Table 3. p-value of post-hoc test for query scores
note. * means $p < .05$; ** means $p < .01$

	Q1	Q2	Q3	Q4	Q5	Q6
Q1	-	1.000	.219	* .010	** .003	**< .001
Q2		-	** .003	** .001	** .002	**< .001
Q3			-	1.000	1.000	**< .001
Q4				-	1.000	**< .001
Q5					-	**< .001
Q6						-

DISCUSSIONS

The current study explored impacts of two SA evaluation methods (SAGAT and SPAM) on participants' workload with respect to future air traffic situations. Simple air traffic control simulations were designed considering traffic density between low to high, and participants' query scores and workload were examined. The result shows participants' workload in real time increased with the experience of simulation regardless of SA evaluation method type. The workload assessed by NASA-TLX suggests that SAGAT imposes higher levels of workload to participants than SPAM. Query score results indicate SPAM is a more predictive method to evaluate participants' SA than SAGAT. Overall, statistical analyses highlight SPAM relatively well predicted in SA evaluation and reducing workload compared to SAGAT.

The result from ANOVA for participants' workload showed that a long experience of simulation leads increased workload for participants. Because the air traffic density increased with the time in the current simulation, the increase imposed high levels of workload for the participants. Further, we manipulated three points where participants were asked to answer SWAT right after SA query presentation (S2, S4, and S5). SWAT scores at the manipulated points were higher than scores at non-manipulated points for either SAGAT or SPAM. This result regarding manipulated points indicates that the SA evaluation imposes high levels of workload for participants regardless of method types. It was hypothesized that SPAM would have a significantly higher workload than SAGAT in such situation, however, no statistical difference in the two methods was found. One possible interpretation for this finding is that working memory affected their workload in the case of SAGAT. SAGAT requires to stop the simulation during answering question whilst participants could be able to answer SPAM with real time simulation. Working memory can be affected in the period from when the simulation is resumed to when the participant regains SA (Sarter & Woods, 1991). In this period, an extra workload might be loaded to the participants because they should make effort to understand and predict past, present and future situations. Additionally, individual differences perhaps influenced participants' attitudes towards two methods (Rose et al., 2019). Further study should discuss the issue of SA evaluation technique with the working memory problem.

As expected, the result implies that SPAM is a relatively more predictive technique compared to SAGAT. In addition, increased air traffic density led decreased SA in both methods. This result can be interpreted with the results of NASA-TLX. The NASA-TLX score which obtained after each simulation also differ between two techniques. Here, SAGAT had a larger influence on workload throughout the simulation than SPAM. Such trend was observed in items of Frustration, Performance, and Effort. According to the interview with participants after the whole experiment, some participants reported that "I did not know that the simulation will stop. So, I could not be aware of the necessity that I should always remember the situations," and "I did not know whether my answer is correct or not." It can be considered that the necessity to remember situations and to answer the queries by relying on only memory shaped participants' feelings for Effort and Performance respectively in the case of SAGAT. This could also influence their feelings for Frustration as a stress. This finding indicates that the higher levels of workload, the more degraded SA levels for the ATC simulation.

Aforementioned findings should be considered with several limitations. First, more sophisticated simulations resembling real air traffic control environment are necessary to predict air controllers' SA. This study programmed a simplified simulation which did not encompass decision-making tasks. Operators' SA during simulations like the original ATC management should be further investigated. Second, in related vein, the current result may not necessarily generalize to controllers' performance and SA. All student participants undertook short practices in terms of SA evaluation methods, and adequate time is needed to be used to all tasks for the participants. Future study could be designed regarding the issue. In spite of these limitations, the current study offered a valuable suggestion into debates about two SA evaluation techniques to solve human factors problems for future air traffic controllers. Future study should investigate the maximum amount of air traffic that the controller are able to maintain high levels of SA as well as to handle situations at the same time regarding for future air traffic condition with original air traffic control tasks.

ACKNOWLEDGMENT

We would like to thank people of Electric Navigation Research Institute for useful discussions and information about CARATS Open Data.

REFERENCES

Athènes, S., Averty, P., Puechmorel, S., Delahaye, D. & Collet, C. (2002). ATC complexity and controller workload: Trying to bridge the gap. *International conference on human-computer interaction in aeronautics 2002*, 56–60

de Winter, J., Esima, Y., Cabrall, C., Hancock, P., & Stanton, N. A. (2019). Situation awareness based on eye movements in relation to the task environment. *Cognition, Technology & Work, 21*, 99–111

Davis, C. G., Danaher, J. W., & Fischl, M. A. (1963). *The influence of selected sector characteristics upon ARTCC controller activities* (Contract No. FAA/BRD-301). Arlington, VA: Matrix

Durso, F. T. & Dattel, A. R. (2004). SPAM: The real-time assessment of SA. *In S. Banbury & S. Trembley, A cognitive approach to situation awareness: Theory, measures and application*, 137–154

Edwards, T., Homola, J., Mercer, J., & Claudatos, L., (2017). Multifactor interactions and the air traffic controller: the interaction of situation awareness and workload in association with automation. *Cognition, Technology & Work*, 19, 687–698

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). *IEEE National Aerospace and Electronics Conference*, 3, 789–795

Endsley, M. R. (1995a). A Taxonomy of Situation Awareness Errors. *In R. Fuller, N. Johnston, and N. McDonald (Eds.), Human Factors in Aviation Operations*, 287–292, Aldershot, England: Avebury Aviation, Ashgate Publishing Ltd.

Endsley, M. R. (1995b). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37, 32–64

Endsley, M. R. (2019). A Systematic Review and Meta-Analysis of Direct Objective Measures of Situation Awareness: A Comparison of SAGAT and SPAM. *Human Factors*

Endsley, M. R., & Garland, D. J. (2000). *Situation awareness analysis and measurement*, Lawrence Erlbaum

Hall, J. & Phelps, R. H. (1983). Nonlaboratory techniques for the study of cognitive and decision-making processes. *A description and selected bibliography* (No. 83–45)

Hart, S. G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, 139–183

Hogg, D.N, Folleso, K., Strand-Volden, F., & Torralba, B. (1995). Development of a situation awareness measure to evaluate advanced alarm systems in nuclear power plant control rooms. *Ergonomics*, 38(11), 2394–2413

International Civil Aviation Organization (ICAO) (1984). Safety Management Manual (*International Civil Aviation Organization website* https://www.unitingaviation.com/publications/safetymanagementimplementation/content/#/)

Kaber, D. B. & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, 5(2), 113–153

Kraemer, J., & Süß, H. M. (2015). Real time validation of online situation awareness questionnaires in simulated approach air traffic control. *Procedia Manufacturing*, 3, 3152–3159

Ministry of Land, Infrastructure, Transport and Tourism (2010). Collaborative Actions for Renovation of Air Traffic Systems (*Ministry of Land, Infrastructure, Transport and Tourism website* https://www.mlit.go.jp/common/000123890.pdf) (in Japanese)

Miyake, S. & Kumashiro, M. (1993). Subjective mental workload assessment Technique: An introduction to NASA-TLX and SWAT and a proposal of simple scoring methods. *The Japanese journal of ergonomics*, 9, 399–408 (Japanese)

Morgan, C. A., Chiappe, D. L., Kraut, J., Strybel, T. Z. & Vu, K. P. L. (2012). An investigation into the potential sources of interference in SA probe techniques. *Advances in human aspects of aviation*, 623–631

Pierce, R. S. (2012). The effect of SPAM administration during a dynamic simulation. *Human Factors*, 54, 838–848

Pritchett, A. R., Hansman, R. J., & Johnson, E. N. (1995). Use of testable responses for performance- based measurement of situation awareness. *Experimental analysis and measurement of situation awareness*, 75–81

Reid, G. B. & Nygren, T. E. (1988). The subjective workload assessment technique: a scaling procedure for measuring mental workload. *Human Mental Workload*, 185–219

Rose, J., Bearman, C., Naweed, A., & Dorrian, J. (2019). Proceed with caution: Using verbal protocol analysis to measure situation awareness. *Ergonomics*, 62, 115–127

Sarter, N. B. & Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology*, 1, 45–57

Updegrove, J. A. & Jafer, S. (2017). Optimization of Air Traffic Control Training at the Federal Aviation Administration Academy. *Aerospace* 2017, 4(4), 50