

Towards the Development of an Automated, Real-Time, Objective Measure of Situation Awareness for Pilots

Sandro Scielzo
L3Harris Technologies
Arlington, Texas
sandro.scielzo@L3Harris.com

Justin C. Wilson, Eric C. Larson
Southern Methodist University
Dallas, Texas
wilsonj@smu.edu, eclarson@smu.edu

ABSTRACT

Measuring mission-critical, higher-order cognitive constructs automatically is a priority within the DoD to achieve third-offset goals, accelerate training of complex skills, and support multi-domain warfare. Situation Awareness (SA) is one such construct, but its measurement is burdensome, relies on post-hoc analyses, and provides little immediate value in training or operational environments. Wilson and colleagues (2020) demonstrated that targeted application of Machine Learning (ML) on biometric data can yield real-time, accurate performance classifiers evaluating pilot eye scan techniques and mental workload. The current study established the relationship between the three levels of SA (i.e., perception, comprehension, and projection) and corresponding proxy ML classifiers, such as visual scan accuracy and mental workload, to lay the foundation for a real-time SA index that is diagnostic of performance. Forty participants, including pilots with varying levels of expertise and an ab initio control group, participated in a meticulously controlled, within-subjects experiment that involved flying an intercept mission using an F/A-18 mixed-reality trainer. Situation awareness was measured using the Situation Awareness Global Assessment Tool (SAGAT), and we used the NASA Task Load Index (NASA-TLX) to gauge mental workload – both subjective metrics are industry gold standards. Visual scan accuracy and mission performance were graded by experienced instructor pilots. Statistical analyses describe the relationship between level of expertise, visual scan accuracy, mental workload, and performance across SA levels. This experiment is unique in that it presents results supporting the validity of our claims while providing a foundation for a real time, objective, and accurate SA index. Findings are discussed in terms of expected benefits for both training complex skills and high-stakes dynamic operational environments.

ABOUT THE AUTHORS

Sandro Scielzo is a multidisciplinary scientist at L3Harris' Military Training sector in the areas of human systems engineering, human factors, and instructional system design. Dr. Scielzo received his PhD in Applied Experimental Human Factors and M.S. in Modeling & Simulation from the University of Central Florida in 2008 and 2005 respectively. His research has concentrated on the validation and implementation of next generation training solutions for military and commercial applications to accelerate learning, guarantee proficiency, transfer knowledge, and maintain skills in the operational environment. Over the course of his career, Dr. Scielzo oversaw a wide portfolio of DoD R&D applied research projects to enhance warfighter training and operational readiness.

Justin Wilson is an experimental test combat systems officer in the United States Air Force. Major Wilson has flown ~1300 hours on 31 different aircraft. He is a USAF Test Pilot School graduate and holds a PhD in Computer Science from the Southern Methodist University. His research focus includes ubiquitous computing, machine learning, and cognitive and context-aware systems.

Eric Larson is an Associate Professor at the Southern Methodist University's Lyle School of Engineering. Dr. Larson has a PhD in Electrical Engineering from Washington University. Dr. Larson holds 53 patents and papers on Machine Learning (ML) with human subjects. His current focus is on incorporating biometric ML for flight test scenarios in M&S and Full-Scale warfighter aircrafts.

Towards the Development of an Automated, Real-Time, Objective Measure of Situation Awareness for Pilots

Sandro Scielzo
L3Harris Technologies
Arlington, Texas
sandro.scielzo@L3Harris.com

Justin C. Wilson, Eric C. Larson
Southern Methodist University
Dallas, Texas
wilsonj@smu.edu, eclarson@smu.edu

INTRODUCTION

Endeavoring to measure warfighters' psychophysiological states and cognitive functions that are diagnostic of performance has been a major aim for the military, industry, and academia alike to develop better training systems which enhance readiness, and effective decision support tools that save lives and improve mission outcomes. Actionable and timely insights based on these human factors are rapidly becoming even more critical, in part due to pivoting to a US third offset strategy (3OS) aimed to fundamentally change how future mission capabilities and corresponding technologies are used to generate battle winning outcomes (Louth, & Taylor, 2016). A specific example is the concept of "Mosaic Warfare", coined by the Defense Advanced Research Projects Agency (DARPA). In Mosaic Warfare individual warfighting platforms are dynamically reconfigured to make a force package that responds to complex near-peer threats (DARPA, 2020). Key to this future warfare paradigm is the need to measure mission-critical, higher-order cognitive constructs automatically to accelerate training of complex skills needed to yield high-level, complex multi-domain mission readiness. Situation Awareness (SA) is one such critical cognitive construct that is closely related to mission performance in complex and dynamic domains. Thus, developing a real-time, automated, and objective measure of SA would significantly improve training for readiness and provide critical actionable information to support current and future warfare operations.

Measuring Higher-Order Cognitive Constructs

Measuring higher-order cognitive constructs such as SA is far from straightforward. It typically requires trained scientists and other resources to design and administer assessment tools using precise methodologies. These data gathering tools must be used in a controlled environment, and the resulting data only offers diagnostic value after post-hoc analyses that vary in duration but are far from immediate. This hard problem prevents real-time use of diagnostic information in either training or operational environments. Solving this problem could radically accelerate achieving readiness in current and future warfare paradigms. The desire to solve this problem is well captured by the Department of Defense's (DoD) Science and Technology (S&T) Communities of Interest (CoIs), which provides advice to senior-level stakeholders in the Department and Congress as a mean to focus future Research & Development (R&D) funding. In fact, DoD S&T CoIs provided a thorough assessment of the need for real-time measurement and prediction of warfighter performance to support current and future technologies that are highly diagnostic of training effectiveness and operational performance (Reliance 21, 2018).

An Emerging Approach to Automated Assessment of Cognitive Constructs

A promising solution takes advantage of advances in biometric sensor capabilities, which enable for the relatively unobtrusive collection of a multitude of physiological indicators (e.g., electrodermal activity, heart rate, pupil dilation) that may be symptomatic of internal states, such as workload, stress, and fatigue. These capabilities are leading to a variety of real-time measurement frameworks such as workload assessment via pupillometry data (e.g., Rafiqi, Nair, & Fernandez, 2014; Wangwiwattana, Ding, & Larson, 2018), or using multiple physiological indicators to diagnose stress, engagement, and fine motor control (Wilson, 2018). Another promising venue is the use of Machine Learning (ML) classifiers that can provide diagnostic information based on multiple and simultaneous stream of biometric data (e.g., Wilson, Nair, Scielzo, & Larson, 2020). However, the application of biometric sensor data and ML to measure a construct such as SA requires a systematic approach to understand how to operationalize the multidimensionality of such a complex construct. What is needed is a validated real-time model of SA. This paper offers an important first step towards that aim by providing empirical evidence that relates performance and biometric measures, such a scan pattern and cognitive load, with measures of SA in an air intercept mission.

BACKGROUND

Measuring SA in real-time requires an approach that considers how it is defined, which key components of SA impact performance, and how well ML classifiers based on biometric data can be developed and validated to map to those key components. This relatively simple approach provides a path forward to the operationalization of SA components necessary to eventually develop a real-time model of SA.

A widely used theoretical model of SA was introduced by Endsley in her seminal article over three decades ago, defining SA as “the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future” (Endsley, 1995, p.36). Thus, three components are at the core of SA formation: perception (Level 1 SA), comprehension (Level 2 SA), and projection (Level 3 SA). In most dynamic, high-stakes domain, having accurate SA is critical for achieving high performance (e.g., Endsley, 1999; Endsley & Garland, 2000), with level of expertise being a major individual mediating factor (Cak, Say, & Misirlisoy, 2020; Endsley, 2006). To date, the only objective measure of SA is the Situation Awareness Global Assessment Technique (SAGAT), which directly compares individuals’ SA (measured via “SA queries” along the three levels of SA) to objective “ground truth” reality (Endsley, 1995). Over the years, SAGAT has shown to have strong sensitivity and validity (e.g., Endsley, 2017). However, as indicated earlier, such measures required time-consuming post-hoc analyses to provide diagnostic information. Thus, there is need for a real-time SA model.

Using Machine Learning Classifiers as SA Component Proxy Metrics

Technological innovations, widespread availability of biometric sensors, and the use of ML methods to train performance classifiers is leading to a burgeoning multidisciplinary field of cognitive and context-aware computing (e.g., Rafiqi et al., 2014; Wangwiwattana et al., 2018). In conjunction with this field of inquiry, recent work by Wilson and colleagues (2020, in submission) proved the viability of using multivariate biometric data to train an automatic classification model using deep convolutional neural networks to classify the quality of pilots’ gaze patterns. When compared to experienced instructors, the classification of scan pattern was identical for 93% of all trials, with instructors agreeing with the model more often than one another. A mental workload classifier is also currently being validated and would similarly allow real time inference of pilot workload. As a result, real-time ML classifiers provide an opportunity to operationalize certain components of SA if they can successfully relate to existing validated measures. These ML classifiers or metrics would effectively become SA component proxy metrics.

Present Study

The empirical study presented in this paper sought to establish the relationship between the three levels of SA (i.e., perception, comprehension, and projection) and ML metrics using military aviation as its domain. Due to the unavailability of ML metrics, analogous measures of visual scan accuracy and mental workload (defined in the next section) were used to evaluate the upper bound of corresponding ML proxy metrics (i.e., gaze and workload classifiers). Table 1 presents main hypotheses between SA levels, ML metrics, performance, and experience. Overall, a critical element of this study was to verify (a) the relationship between SA, workload, performance, and experience, as hypothesized from decades of research (additional pertinent hypotheses are presented in the results section of this paper), and (b) the relationship between SA levels and corresponding ML metrics, assessed by analogous measures of visual scan accuracy and mental workload.

Table 1. Main Study hypotheses between SA levels, ML metrics, performance, and experience

	SA Level 1 (Perception of elements in current situation)	SA Level 2 (Comprehension of current situation)	SA Level 3 (Projection of future states)
Hypotheses*			
ML Metric Hypotheses	Level 1 SA is related to visual scan accuracy and mental workload	Level 2 SA is related primarily to mental workload	Level 3 SA is related to mental workload
Performance Hypotheses	Better Level 1 SA leads to better performance	Better Level 2 SA leads to better performance	Better Level 3 SA leads to better performance
Experience Hypotheses	The higher the expertise, the higher the Level 1 SA	The higher the expertise, the higher the Level 2 SA	The higher the expertise, the higher the Level 3 SA

* All hypothesis statements imply expected significant correlations or mean differences

METHOD

Participants and Study Design

Participants. A total of 40 subjects participated in this study: twenty pilots, nine operators, and eleven novices. Pilots had varied military flying experience in fast jet, wide-body, or rotary aircrafts. Operators had flight experience as naval flight officers (NFOs), combat systems officers (CSOs), remotely piloted aircraft (RPA) sensor operators, or avionics technicians. Novices had no aircraft flying experience.

Design. A 2 Event (air intercept mission event 1 & 2) x 4 Expertise (levels 1 through 4) mixed-model design was used, with Event as the within-subjects independent variable (each participant flew the air intercept mission twice), and Expertise as the between-subjects independent variable. Table 2 presents expertise levels and definitions.

Table 2. Expertise levels and definitions

Level	1 (Pilot – Congruent)	2 (Pilot – other)	3 (Operator)	4 (Novice)
Definition	Pilots with F/A-18 experience	Pilots without F/A-18 experience	Operators with some flight experience	Individuals with no flight experience

Materials, Apparatus, and Procedures

Simulation Device. Participants used the L3Harris Blue Boxer™ Extended Reality (BBXR) deployable training system (Hanson, 2018). The Blue Boxer is a low-footprint mixed reality training device that combines physical cockpit attributes (accurate front main panel and stick & throttle with high-precision hand tracking) with a Virtual Reality (VR) environment (see Figure 1). The environment was displayed via the HTC VIVE Pro Eye VR headset with integrated Tobii eye tracker, which can be worn with eyeglasses. The Blue Boxer™ was selected because it maintains a high refresh rate of 90Hz, found to negate VR induced sickness (of the 40 participants, no sickness was reported), and because of having full control over the visual environment.



Figure 1. L3Harris F/A-18 Blue Boxer™ Mixed Reality Device (Hanson, 2018)

Air Intercept Mission. This scenario was designed to offer some challenge to experienced pilots, but feasible and engaging for novices. Figure 2 illustrates the intercept profile (not to scale), consisting of a F/A-18 blue air (flown by the participants) versus a MiG 29 red air (flown by a confederate using a lower fidelity man-in-the-loop system; the confederate maintained constant speed and altitude throughout the scenario). The two aircraft started head-to-head, thirty nautical miles apart, at the same speed, but different altitudes. The intercept mission is composed of the following tasks: (1) obtain a radar lock on bogey aircraft, (2) take 30° offset left or right, (3) descend 10° nose low to the bogey's altitude or slightly below, (4) level-off and accelerate to > 400 knots calibrated airspeed (KCAS), (5) execute an intercept/escort profile, and (6) close for visual identification (VID) and verification of the aircraft markings (fin flash).

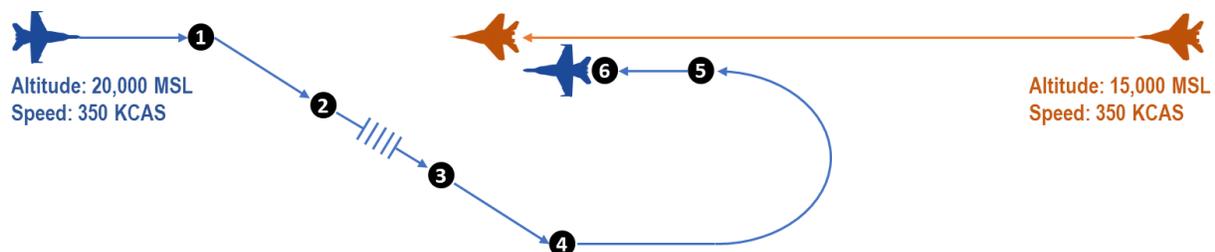


Figure 2. Air Intercept Mission Profile and Initial Conditions

SAGAT Measure. A total of 27 SA queries were developed across perception, comprehension, and projection levels to assess participants’ SA accuracy. These queries were developed using Endsley’s (1988) methodology and divided into two SAGAT “stops” (i.e., queries administered to participants after freezing the simulation), with 14 queries in stop “Alpha”, 13 in stop “Bravo”, and designed to take about two minutes to complete. Table 3 provides examples of SA queries by level of SA. The scoring method involved two Instructor Pilot (IP) SMEs determining ground truth for each query by reviewing air intercept mission recordings for each participant. Specific criteria and thresholds were applied to score each query. SA was scored at three levels: good, acceptable, and unacceptable.

Table 3. Total SA queries by SA level with examples

SA	Measure of	N	Examples SA Queries
Level 1	<i>Perception of elements in current situation</i>	10	What is your separation distance? What is your airspeed?
Level 2	<i>Comprehension of current situation</i>	12	What is the impact of the closure rate? Given your current bank angle, what is the impact on the intercept?
Level 3	<i>Projection of future states</i>	5	What will be your rollout intercept positioning? When will you be within 1 mile?

NASA Task Load Index. The NASA task load index (NASA-TLX) developed by Hart and Staveland (1988) is the gold-standard for assessing workload and was used as a proxy for a mental workload classifier. A variation referred to as the Raw NASA-TLX (RTLX) was used to measure workload, which eliminates the subscales weighting process thereby making the overall scale much simpler to administer. RTLX sensitivity compared to the traditional NASA-TLX was found to vary greatly, ranging from lower to higher sensitivities (Hart, 2006). In this study RTLX was adopted to minimize experimental disruption and mitigate time constraints, and ratings across NASA-TLX dimensions were averaged to create an estimate of overall workload. For consistency, the term NASA-TLX is used throughout the paper when referring to participants’ workload.

Task Performance and Visual Scan Accuracy Measures. Overall Measure of Performance (MOP) on the air intercept mission was assessed by the same two IP SMEs with objective criteria for each of the mission task presented in Table 4 (note that only simple criteria are shown. SMEs used additional complex criteria). Task criteria were scored on a 3-point scale (green = good, yellow = acceptable, red = unacceptable). Visual scan accuracy was used as a proxy for the ML gaze classifier, and also assessed by two IP SMEs. Each SME reviewed mission videos which tracked participants’ gaze convergence (see Figure 3) and scored visual scan accuracy for the overall mission on a 3-point scale: efficient scan, acceptable scan, and unacceptable scan. Scoring heuristic for visual scan assessment is: (1) efficient, with scan accuracy ≥ 90% of time and no overshoot, (2) acceptable, with scan accuracy ≥ 70% of time and no overshoot (or 90% with overshoot), and (3) unacceptable, with scan accuracy < 70% of time or overshoot (i.e., missing the intercept).

Table 4. Measure of Performance (MOP) Criteria

MOP	Criteria		
	“Green”	“Yellow”	“Red”
1. Radar lock	≥20 NM	≥17 NM	≤17 NM
2. Take Offset	≥27°	≥20°	≤20°
3. Descend	<15000’ by 10NM	<15500’ by 10NM	>15500’ by 10NM
4. Accelerate	≥450 KTS at passing bogey alt.	≥420 KTS at passing bogey alt.	<420 KTS at passing bogey alt.
5. Intercept	<Heading 1NM sep.	<Heading 2NM sep.	<Heading 3NM sep.
6. VID	<0.5NM, V _c <100KTS	<0.5NM, V _c ≤50KTS	<0.5NM, V _c >50KTS

Note: NM = nautical mile, V_c = velocity closure, KTS = knots



Figure 3. Screenshots showing gaze convergence

Procedures. All experimental procedures were approved by a University’s institutional review board. After obtaining verbal consent, participants were asked to sit in the Blue Boxer™. While seated, participants were briefed on the air intercept mission reviewing each task and expected performance criteria. Participants then wore the HTC VIVE Pro Eye VR headset, and after ensuring good fit and adjusting inter-pupil distance, the automated eye tracking calibration

was performed. A practice session followed, allowing participants to fly the F/A-18 and familiarize themselves with aircraft's controls and overall VR environment. Novices were given more time to practice and to rehearse the air intercept mission once while being guided by the experimenter. Once all questions were answered, participants proceeded with the air intercept mission event 1. About one-third of the way through the mission, the simulation was frozen and the visual environment blacked-out to prevent participants from reading information, at which point the SAGAT stop Alpha was administered verbally by the experimenter, who recorded each answer. Participants then proceeded with the mission, and at about two-thirds of the way through the mission, SAGAT stop Bravo was administered. After answering the SA queries, participants completed the mission. At the end of the mission, participants completed the NASA-TLX, which was also administered verbally. The same procedures were repeated for the air intercept mission event 2 (identical mission). Once both events were completed, participants were debriefed.

RESULTS

Data loss for 3 out of 40 original participants occurred due to system recording errors. Thus, the final sample used for all analyses reported in this section is 37. Specifically, across the four levels of expertise sample sizes were: 8 pilots (congruent), 11 pilots (other), 9 operators, and 9 novices. Data scoring on the SAGAT, MOP, and visual scan accuracy measures was conducted by two IP SMEs. The two SMEs followed a two-step process for scoring: independent scoring, followed by a joint session to resolve disagreements. The outcome of this process yielded 98.25% agreement for SAGAT and 98.46% agreement for MOP and visual scan accuracy.

Normality Assessment and Psychometric Properties of Subjective Dependent Measures

The data, based on a total of 37 participants, was screened for normality on all dependent measures of SA, mental workload, performance, and visual scan accuracy (event 1 & 2 measures, and aggregates). Table 5 presents the outcome of this analysis by indicating with a "Y" when a measure was significantly skewed or kurtotic. Of the twenty-one dependent measures, six were not normally distributed. Specifically, one SAGAT measure was significantly negatively skewed and leptokurtotic. All NASA-TLX measures were normally distributed, but performance measures were all significantly negatively skewed indicating that most participants tended to perform well. Finally, visual scan accuracy distributions (other than the aggregate) were found to be platykurtotic. Despite these findings, corrections were not applied to preserve interpretability in the context of the other measures.

Table 5. List of non-normally distributed measures in terms of skewness and kurtosis

Measure	N	Skewness	Std. Error of Skewness	Sig.*	Kurtosis	Std. Error of Kurtosis	Sig.*
SAGAT event 2 (Level 2 SA queries)	37	-0.93	0.39	Y	1.74	0.76	Y
MOP aggregate event 1	36	-0.86	0.39	Y	0.31	0.77	
MOP aggregate event 2	37	-1.69	0.39	Y	5.61	0.76	
MOP total aggregate	37	-0.85	0.39	Y	1.38	0.76	
Visual scan accuracy event 1	35	0.41	0.40		-1.57	0.78	Y
Visual scan accuracy event 2	37	0.27	0.39		-1.60	0.76	Y

* Significance, as indicated by Y, is determined by dividing the absolute value of skewness and kurtosis by their respective standard error and verifying that the obtained value > 1.96.

Psychometric properties of the two gold standard scales used in this study – SAGAT, and NASA-TLX – were also investigated. Table 6 provides information on the reliability of both SAGAT and NASA-TLX subjective scales for both testing events. Specifically, Cronbach's *alpha* coefficient is provided as an estimate of internal consistency reliability. Standards for interpreting the *alpha* coefficient vary, but a common heuristic for multidimensional scales such as SAGAT and NASA-TLX requires a coefficient greater than 0.5 to be acceptable, while a minimum value 0.65 is typically recommended. Thus, both the scales indicate they have acceptable to excellent internal reliability.

Table 6. Internal reliability of SAGAT and NASA-TLX measures

SAGAT Measure	N	<i>alpha</i>	NASA-TLX Measure	N (items)	<i>alpha</i>
SAGAT event 1	26	0.76	NASA-TLX event 1	6	0.89
SAGAT event 2	26	0.66	NASA-TLX event 2	6	0.93

SA Relationship with Eye Scan Accuracy, Mental Workload, and Performance

Table 7 presents relationship strength between SA, eye scan accuracy, mental workload, and performance measures using Pearson's correlation coefficient r . The hypothesis that SA levels of perception (level 1) would relate to visual scan accuracy while comprehension (level 2) and projection (level 3) would not was partly supported, indicating that the ability to build proper level 1 and 2 SA is greatly associated to visual scan accuracy. Across both test events, overall SA (all levels) and visual scan accuracy were also found to be significantly correlated, $r(35) = 0.48, p = .001$. Next, the hypothesis that SA levels and mental workload would negatively relate was partly met, showing that while level 1 and level 2 SA did significantly negatively correlate with NASA-TLX for each test event, level 3 SA did not. At the aggregate level, overall SA and mental workload were found to be significantly negatively correlated, $r(35) = -0.56, p = .000$. These results indicate that while better SA is associated with lower mental workload, a pilot's ability to project future states (level 3 SA) is not related to mental workload. Finally, the hypothesis that SA levels would correlate with performance was partly met, with significant correlations at the aggregate level for level 2 and 3 SA. Although approaching significance, level 1 SA was not significantly correlated with performance. Figure 4 shows these significant relationships with scatterplots and trendlines.

Table 7. Correlations of SA measures with visual scan accuracy, NASA-TLX, and performance

Measure	Visual Scan		NASA-TLX		MOP	
	r	p	r	p	r	p
Event 1 SA measures						
SAGAT - Level 1 SA queries	.41	.01**	-.49	.00**	.13	.22
SAGAT - Level 2 SA queries	.40	.01**	-.29	.04*	.21	.11
SAGAT - Level 3 SA queries	.10	.29	-.05	.38	.18	.15
SAGAT aggregate	.23	.09	-.41	.01**	.23	.09
Event 2 SA measures						
SAGAT - Level 1 SA queries	.41	.01**	-.58	.00**	.20	.12
SAGAT - Level 2 SA queries	.36	.02*	-.54	.00**	.16	.17
SAGAT - Level 3 SA queries	.20	.09	-.20	.13	.25	.07
SAGAT aggregate	.25	.07	-.67	.00**	.25	.07
Aggregate SA measures (across events)						
SAGAT - Level 1 SA queries	.42	.01**	-.54	.00**	.24	.07
SAGAT - Level 2 SA queries	.50	.00**	-.50	.00**	.44	.01**
SAGAT - Level 3 SA queries	.13	.23	-.15	.19	.31	.03*
SAGAT total aggregate	.48	.00**	-.56	.00**	.40	.01**

* Correlation significant at the 0.05 level (1-tailed); ** Correlation significant at the 0.01 level (1-tailed)

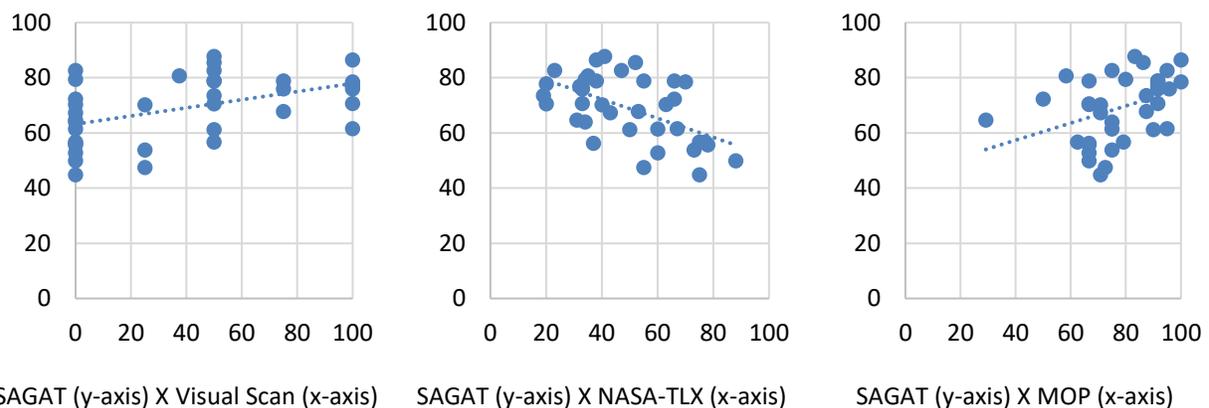


Figure 4. Scatterplots with trendlines for SA with Visual Scan Accuracy, Cognitive Load, and Performance

Having established the relationship between SA and main dependent measures, we partly verified the hypothesis that both visual scan accuracy and mental workload would correlate with performance. In fact, Table 8 shows that only visual scan accuracy was significantly correlated with performance. Negative correlations between NASA-TLX and performance were found but were not significant.

Table 8. Correlations of performance measures with visual scan accuracy and NASA-TLX

Measure	Visual Scan		NASA-TLX	
	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>
MOP aggregate event 1	.65	.00**	-.14	.21
MOP aggregate event 2	.76	.00**	-.20	.13
MOP total aggregate	.76	.00**	-.15	.19

* Correlation significant at the 0.05 level (1-tailed); ** Correlation significant at the 0.01 level (1-tailed)

Impact of Experience across Dependent Measures

Impact of Experience on SA. Analyses were performed using SPSS General Linear Model Multivariate Analyses of Variance (MANOVA), and the results are reported with $\alpha = 0.05$, one-tailed, due to our stated directional hypotheses. Besides unequal cell sizes, parametric assumptions of linearity and homogeneity of variance were met. As Hypothesized, there was a statistically significant difference for measures of SA (level 1, level 2, level 3, and SAGAT aggregate) based on level of expertise (1 through 4), $F(12, 79.66) = 2.28$, $p = 0.015$; Wilk's $\Lambda = 0.46$, partial $\eta^2 = 0.23$. Table 9 lists between-subjects effects for each measure of SA, showing that measures other than level 3 SA (projection) had significant differences based on level of expertise. Table 10 lists significant post hoc tests using the Bonferroni correction and identifies which mean comparisons were statistically different. Figure 5 plots SA accuracy (level 1, level 2, and aggregate SA) across levels of experience.

Table 9. Between-subjects effects for each measure of SA

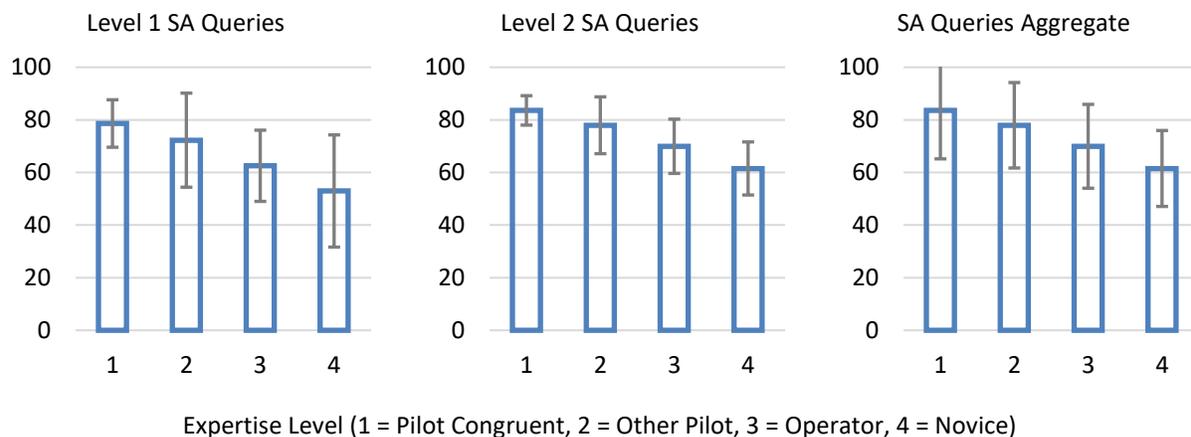
Dependent Variable	Sum of Squares	<i>df</i>	Mean Square	<i>F</i>	<i>p</i>	partial η^2
SAGAT - Level 1 SA queries	3332.556	3	1110.852	4.131	.007**	.273
SAGAT - Level 2 SA queries	2435.177	3	811.726	8.773	.000**	.444
SAGAT - Level 3 SA queries	1495.270	3	498.423	1.887	.076	.146
SAGAT aggregate	2157.417	3	719.139	8.433	.000**	.434

* Significant at the 0.05 level; ** significant at the 0.01 level

Table 10. Significant post hoc tests comparisons between expertise levels and SA measures

Dependent Variable	Levels Compared	M_1	SD_1	M_2	SD_2	<i>p</i>
SAGAT - Level 1 SA queries	1 and 4	78.62	9.02	52.96	21.33	0.009**
	2 and 4	72.27	17.89	52.96	21.33	0.040*
SAGAT - Level 2 SA queries	1 and 3	83.60	5.60	69.96	10.32	0.019*
	1 and 4	83.60	5.60	61.52	10.09	0.000**
SAGAT aggregate	2 and 4	77.94	10.80	61.52	10.09	0.002**
	1 and 3	69.60	18.43	52.14	15.94	0.009**
SAGAT aggregate	1 and 4	69.60	18.43	60.77	14.44	0.000**
	2 and 4	65.63	16.24	60.77	14.44	0.003**

* Significant at the 0.05 level; ** significant at the 0.01 level

**Figure 5. SA Means and Standard Deviations (Level 1, Level 2, and Aggregate SA) by Level of Expertise**

Impact of Experience on Visual Scan Accuracy, Cognitive Workload, and Performance. Analyses were performed using SPSS General Linear Model Univariate Analyses of Variance (ANOVA), and the results are reported with $\alpha = 0.05$, one-tailed, due to our stated directional hypotheses. Besides unequal cell sizes and non-normal distributions for certain measures (see Table 5), parametric assumptions of linearity and homogeneity of variance were satisfactory. Table 11 lists between-subjects effects results for event 1 visual scan accuracy, cognitive workload, and performance. Overall, as hypothesized, visual scan accuracy, cognitive workload, and performance had significant differences based on level of expertise (see Table 11). Table 12 shows post hoc tests using the Bonferroni correction and identifies which mean comparisons were statistically different. Figure 6 plots visual scan accuracy, cognitive workload, and performance across levels of experience.

Table 11. Between-subjects effects for visual scan accuracy, cognitive workload, and performance for event 1

Dependent Variable	Sum of Squares	<i>df</i>	Mean Square	<i>F</i>	<i>p</i>	partial η^2
Visual scan accuracy	22792.190	3	7597.397	8.340	.000**	.431
NASA-TLX	4714.690	3	1571.563	6.407	.001**	.368
MOP	1785.618	3	595.206	3.003	.022*	.214

* Significant at the 0.05 level; ** significant at the 0.01 level

Table 12. Significant post hoc tests comparisons between expertise levels for visual scan accuracy, cognitive load, and performance

Dependent Variable	levels compared	<i>M</i> ₁	<i>SD</i> ₁	<i>M</i> ₂	<i>SD</i> ₂	<i>p</i>
Visual scan accuracy	1 and 3	79.69	28.30	25.00	35.36	0.002**
	1 and 4	79.69	28.30	11.11	18.16	0.000**
	2 and 4	47.73	34.38	11.11	18.16	0.033*
NASA-TLX	1 and 3	37.50	16.63	60.00	17.54	0.017*
	1 and 4	37.50	16.63	61.22	15.90	0.012*
	2 and 3	38.45	12.93	60.00	17.54	0.013*
	2 and 4	38.45	12.93	61.22	15.90	0.009**
MOP	1 and 3	85.80	15.25	70.28	19.12	0.090†

* Significant at the 0.05 level; ** significant at the 0.01 level; † trend at the 0.10 level

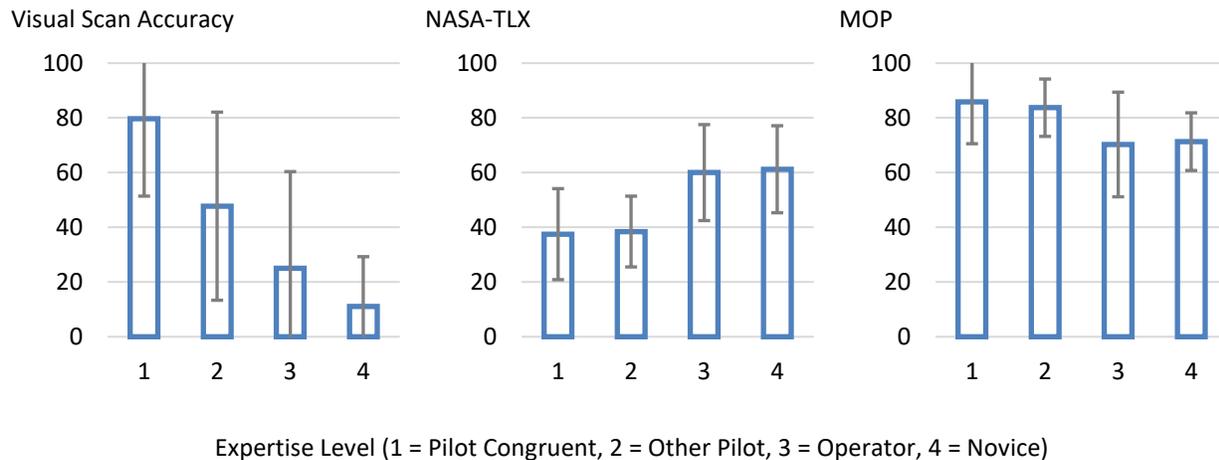


Figure 6. Means and Standard Deviations for visual scan accuracy, cognitive workload, and performance by Level of Expertise

Training Effects across Dependent Measures

SPSS General Linear Model Repeated Measures Analyses were performed on a 2 x 4 mixed model, with event (1 & 2) as the within-subjects independent variable, and expertise level (1 through 4) as the between-subjects independent variable. The model was run on the four main dependent measures of interest: SA, visual scan accuracy, cognitive

load, and performance. Results are reported with $\alpha = 0.05$, two-tailed, due to not having a-priori hypotheses. Besides unequal cell sizes, and non-normal distributions for measures listed in Table 5, parametric assumptions of linearity and homogeneity of variance were satisfactory. Table 13 reports all significant results. Overall, a significant main effect for event was found on SA, indicating that across events SA significantly improved across all experience levels. A peculiar finding is the significant interaction of event by expertise for cognitive workload, indicating that cognitive workload reduced across events for all experience levels other than for novices, for whom it increased. Table 14 presents simple effects analysis for both SA and cognitive load, applying the Bonferroni correction, to pinpoint the nature of those differences. All means and standard deviations are presented in Figure 7.

Table 13. Significant main effects and interactions for SA, visual scan accuracy and performance

Dependent Variable	Sum of Squares	df	Mean Square	F	p	partial η^2
Event Main Effects						
SAGAT aggregate	1218.744	1	1218.744	19.168	.000	.375
Event x Experience Interaction						
NASA-TLX	731.998	3	243.999	3.770	.021	.281

* Significant at the 0.05 level; ** significant at the 0.01 level

Table 14. Simple effects with means and standard deviations for SA and cognitive load measures across expertise levels and events

Dependent Variable	Expertise level	M_{Event1}	SD_{Event1}	M_{Event2}	SD_{Event2}	p
SAGAT aggregate	1	74.24	7.19	83.84	4.39	0.003**
	2	70.25	12.97	77.12	8.63	0.059†
	3	62.81	10.53	67.49	11.79	0.384
	4	50.00	10.12	62.06	10.81	0.016**
NASA-TLX	1	39.86	19.60	34.71	16.97	0.094†
	2	40.36	11.65	36.45	15.32	0.181
	3	66.62	14.90	58.00	21.33	0.061†
	4	61.71	16.97	71.71	14.01	0.180

* Significant at the 0.05 level; ** significant at the 0.01 level; † trend at the 0.10 level

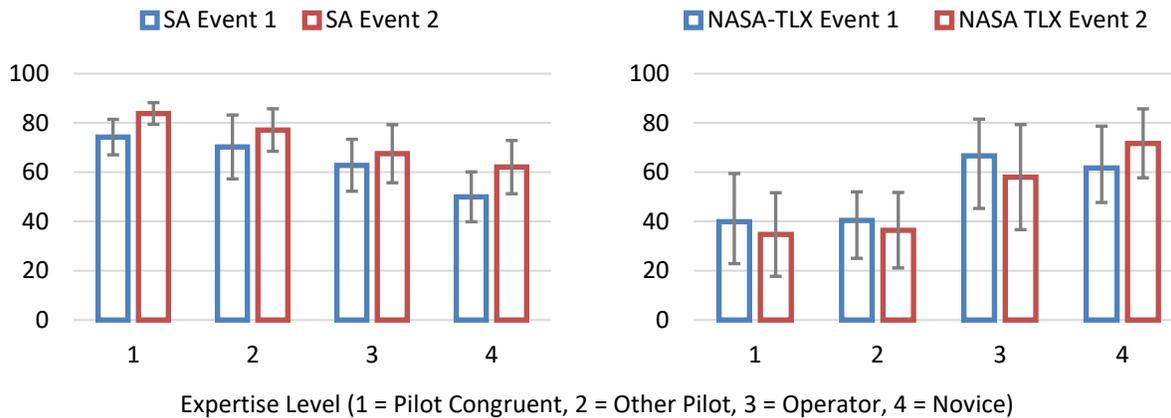


Figure 7. Means and Standard Deviations for visual scan accuracy, cognitive workload, and performance by Level of Expertise

DISCUSSION

The empirical study presented in this paper established the relationship between the three levels of SA (i.e., perception, comprehension, and projection) and analogous ML metrics (i.e., visual scan accuracy and mental workload as proxies for their corresponding ML classifiers). A highly dynamic air intercept mission was adopted as a testbed to test the study hypotheses and to explore training effects. Table 15 presents hypotheses testing outcomes between SA levels, ML metrics, performance, and experience, showing that most hypotheses were supported or partly supported.

Table 15. Hypotheses testing outcomes between SA levels, ML metrics, performance, and experience

Hypotheses	SA Level 1	SA Level 2	SA Level 3
ML Hypotheses	S Level 1 SA is related to visual scan accuracy and mental workload	PS Level 2 SA is related primarily to mental workload	NS Level 3 SA is related to mental workload
Performance Hypotheses	NS Better Level 1 SA leads to better performance	S Better Level 2 SA leads to better performance	S Better Level 3 SA leads to better performance
Experience Hypotheses	S The higher the experience, the higher the Level 1 SA	S The higher the experience, the higher the Level 2 SA	NS The higher the experience, the higher the Level 3 SA

Note: S = hypothesis supported; PS = hypothesis partly supported; NS = hypothesis not supported

Expertise, Training Effects, and Limitations

The impact of expertise on SA revealed that although higher experience led to better SA levels 1 and 2, it surprisingly did not for SA level 3. This finding is difficult to reconcile since projecting future systems states is obviously related to practicing corresponding knowledge, skills, and abilities. It may be that the SA queries used to measure level 3 SA were not sensitive enough, or that experienced pilots did not practice air intercept projections in a while (indicating skill decay). However, differences in overall SA were driven by expertise levels. Moreover, level of expertise was also directly related to visual scan accuracy, workload, and performance.

Training effects were found on overall SA and workload, but not on performance. That is, although SA accuracy improved, and workload decreased across the two events, significant performance improvements were not found. Limited samples across levels of expertise could have led to this outcome. However, these results underline the complex relationship between critical cognitive constructs, training, and performance. In fact, results showed a significant interaction between level of expertise and workload, indicating that workload decreased with additional training for all expertise level other than novices, for whom workload increased. Without additional insights, it is difficult to interpret this interaction. It may be that on the second event novices were better able to grasp the various mission objectives, thereby increasing their workload.

Gaze Accuracy and Mental Workload as Valid Proxies of SA

Visual scan accuracy was indeed found to be strongly related to SA level 1, indicating that a proper visual scan of the environment supports perceiving mission-specific information from that environment. However, visual scan accuracy was also related to SA level 2, indicating that a proper visual scan also supports comprehension of the various perceived elements germane to the mission. This may indicate that although theoretically separate, both SA levels 1 and 2 resource similar cognitive processes associated with visual scan behaviors. A very strong relationship between visual scan accuracy and performance was also found, demonstrating that proper scanning behaviors are essential to positive mission outcomes. As a result, a valid and accurate ML gaze classifier could provide strong, real-time, diagnostic information on scan accuracy, SA perception and comprehension levels, and performance.

Mental workload also related strongly with SA levels 1 and 2, but surprisingly not with SA level 3. This may indicate that both SA level 1 and 2 use working memory resources, while SA level 3 may use more automated, top-down processes to project future system states. Also surprising is the lack of significant relationship between mental workload and performance. This finding may indicate that even the high-level of workload encountered in this study were within the psychophysiological zones of adaptability (see Hancock & Warm, 1989). That is, performance was maintained even with higher levels of workload. This concept is critical in both training and operations. As a result, a ML classifier of mental workload needs to at least be able to classify mental workload as acceptable (within zone of adaptability) or unacceptable (within zone of instability). Overall, these results underline the viability of ML classifiers to potentially operationalize SA. Specifically, the gaze and workload classifiers currently being developed by Wilson and colleagues (2020, in submission) could serve as proxy measures of SA levels 1 and 2. A real-time model of SA would also need to take into consideration level of expertise to yield an accurate SA index.

Implications for Training and Operations

The results presented in this study provide initial justification for using ML classifiers as proxy metrics of SA. Implications for training and operations could be profound. In training environments, real-time SA could identify a breakdown in perception, allowing to remediate poor visual scans, as an example. A breakdown in comprehension

could also be mitigated by ensuring students remain within the zone of maximal adaptability via real-time training complexity adaptation. Thus, mission readiness could be achieved more effectively and rapidly. Similarly, applying real-time SA measurement in operational settings could lead to a better understanding on how to dynamically reconfigure warfighting platforms in a Mosaic Warfare paradigm, or at the very least, help prevent breakdown in decision-making due to lack of SA.

CONCLUSION

Measuring higher-order cognitive constructs objectively and in real-time is rapidly becoming a tangible reality. Future research needs to ascertain the use of actual ML classifiers as valid SA component metrics before a real-time model of SA can be established. Additional research should also be generalizable to other domains, not just aviation, to have a broader impact on multi-domain, joint, training and operations. Achieving this result would significantly evolve the business of training and development of decision-making support tools.

REFERENCES

- Cak, S., Say, B., & Misirlisoy, M. (2020). Effects of working memory, attention, and expertise on pilots' situation awareness. *Cognition, Technology & Work*, 22(1), 85-94.
- DARPA Tiles Together a Vision of Mosaic Warfare. Retrieved May 11, 2020, from <https://www.darpa.mil/work-with-us/darpa-tiles-together-a-vision-of-mosaic-warfare>
- Endsley, M. R. (1997, October). Supporting situation awareness in aviation systems. In *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation* (Vol. 5, pp. 4177-4181). IEEE.
- Endsley, M. R. (1988, May). Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE 1988 national aerospace and electronics conference* (pp. 789-795). IEEE.
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human factors*, 37(1), 32-64.
- Endsley, M. R. (2006). Expertise and situation awareness. *The Cambridge handbook of expertise and expert performance*, 633-651.
- Endsley, M. R. (2017). Direct measurement of situation awareness: Validity and use of SAGAT. In *Situational Awareness* (pp. 129-156). Routledge.
- Endsley, M. R., & Garland, D. J. (2000, July). Pilot situation awareness training in general aviation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 44, No. 11, pp. 357-360). Sage CA: Los Angeles, CA: SAGE Publications.
- Hancock, P. A., & Warm, J. S. (1989). A dynamic model of stress and sustained attention. *Human factors*, 31(5), 519-537.
- Hanson, T. (2018). *L3 Introduces First-Ever High-Fidelity, Mixed Reality Deployable Training Simulator*. Retrieved May 11, 2020, from <https://www.l3t.com/link/press/l3-introduces-first-ever-high-fidelity-mixed-reality-deployable-training-simulator>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.
- Hart, S. G. (2006, October). NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, No. 9, pp. 904-908). Sage CA: Los Angeles, CA: Sage Publications.
- Louth, J., & Taylor, T. (2016). The US Third Offset Strategy: Hegemony and Dependency in the Twenty-First Century. *The RUSI Journal*, 161(3), 66-71.
- Rafiqi, S., Nair, S., & Fernandez, E. (2014, May). Cognitive and context-aware applications. In *Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments* (pp. 1-7).
- Reliance 21 – DoD Communities of Interest, Human Systems (2018). Retrieved May 11, 2020, from https://defenseinnovationmarketplace.dtic.mil/wp-content/uploads/2020/03/Human_Systems_NDIA_2020.pdf
- Wangwuwattana, C., Ding, X., & Larson, E. C. (2018). PupilNet, Measuring Task Evoked Pupillary Response using Commodity RGB Tablet Cameras: Comparison to Mobile, Infrared Gaze Trackers for Inferring Cognitive Load. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4), 1-26.
- Wilson, J. (2018). *Human Performance Optimization Using Biometric Indices: Validation for Targeted Pilot Population*. Unpublished technical report for L3 Technologies, Inc.
- Wilson, J., Nair, S., Scielzo, S., & Larson, E. C. (2020). Automatic Gaze Classification for Aviators: Using Multi-task Convolutional Networks as a Proxy for Flight Instructor Observation. *International Journal of Aviation, Aeronautics, and Aerospace*, 7(3).