Situation Awareness, Workload and Performance: New Directions

Don Harris, Heikki Mansikka and Kai Virtanen

Abstract

The assessment of pilot mental workload and situation awareness (SA) is vital for many aerospace applications, for example the validation and verification of designs; evaluation of tactics, techniques and procedures (TTPs); as a component of aircraft certification, or the assessment of task flows on the flight deck. However, the practical utilisation of these measures poses methodological and measurement challenges. Furthermore, pilots do not fly alone: they are a part of a team. In this chapter measurement techniques developed from well-known, commonly used methods of SA and workload measurement (Situation Awareness Global Assessment Technique – SAGAT, Endsley, 1988; the NASA TLX workload scale, Hart & Staveland, 1988 and physiological measures of workload) are described. These developments address some of the measurement issues and shortcomings posed by these commonly used approaches and are extended to describe team performance. The commonly observed dissociation between measures of workload, SA and performance is addressed and the theoretical basis for these sometimes divergent results is described. The chapter concludes with a model describing an integrated approach to the assessment of SA, including team SA, workload and performance.

Performance, Situation Awareness and Workload

There is an intimate relationship between pilot performance, mental workload and situation awareness (SA), however this association is not straightforward: low workload does not necessarily result in high performance, and it is not inevitably associated with high SA (Mansikka et al., 2019a). It is important to understand the relationship between these three concepts when evaluating any new piece of flight deck equipment or procedure. The assessment of performance alone can potentially be misleading (e.g. Mansikka et al., 2021a, b). Assessment of workload is frequently used in conjunction with measures of SA when comparing between options for new designs or procedures, especially when they produce similar levels of performance. Often the workload question becomes not 'which option produces the best performance' but 'what is the cost in terms of information

processing to achieve a certain level of performance'? Measures of SA provide a means to describe how successfully pilots can acquire and integrate information in a complex flight environment.

Workload is a measure of the cognitive load experienced by a pilot, in contrast to task load, which is the amount of work that interacting with the system actually requires. It can be conceptualised as the information processing 'cost' to perform a given flight task, relating specifically to the finite capacity of cognitive resources (Harris, 2011). However, the level of mental effort invested by the pilot is based upon their subjective assessment of the required performance criteria, *not* the objective task load. Put more simply, from a cognitive standpoint you work as hard as you think that you need to. If pilots are not aware of the actual demands of a task, their workload may be relatively low but ultimately so too will be their performance (Mansikka et al., 2019a). Alternatively, even though the situation does not obviously demand it, a pilot may invest a great deal of cognitive work to attain high SA, hence be under considerably more workload but in the long run, also attain a superior level of performance.

When performing the same task, in the same aircraft and in identical conditions, two pilots can both produce identical, high levels of performance. However, one pilot may experience lower workload compared to the other (Vidulich & Wickens, 1986) and so have more 'spare cognitive capacity' to deal with other issues if required (see region A in Figure 1). If their performance is not limited by other factors, their workload will increase, and performance will eventually degrade when they have no more excess capacity to cope with the increasing task demands (region B). In Figure 1 the vertical bars represent the pilots' overall cognitive capacity, where a darker shading represents more spare capacity and thus lower workload. The solid line represents performance. In region A, Pilot 1 and Pilot 2 can maintain equal levels of performance as they both have enough cognitive capacity for the task. However, Pilot 2 has less spare capacity than Pilot 1. In region B, the increase in task load further taxes the pilots' cognitive capacity such that they no longer can maintain their performance. At a similar task load, Pilot 1 still possesses more cognitive capacity and can thus maintain higher performance than Pilot 2. In region C, both pilots have depleted their cognitive spare capacity, workload (determined by the amount of cognitive spare capacity) is high, and performance is poor, regardless of their efforts. Workload can become dissociated from performance particularly if the task is resource limited (Yeh & Wickens, 1988).

Similarly, the relationship between SA and performance may either be relatively weak (e.g., Fracker, 1991; Endsley, 2019) or complex and unclear (Sulistyawati et al., 2009; Mansikka et al., 2019a). It has already been stated that awareness of task demands will partially determine workload. This awareness is predicated upon SA. Mansikka et al. (2019a) observed that in simulated air combat scenarios, when awareness of the tactical situation was low, pilots exhibited a combination of low

workload and low performance: they were not aware that they should be working harder to attain a higher level of SA. Furthermore, in a highly dynamic, uncertain environment such as air combat, success may occasionally be a product of chance factors and *vice versa* (Mansikka et al., 2021d). Pilots' work is often described as a form of Input-Process-Output (IPO) model, where processes involving both individual and team cognitive activities are converted into outputs. In simple terms, the output is a result of what pilots do and is typically assessed only with output performance measures, irrespective of how they arrived at that output. To understand how the pilots reached their output, measures targeting output performance should be supplemented by adjunct measures such as SA and workload (Mansikka et al., 2021b, c) and other process measures (Mansikka et al., 2021a). Assessment of SA can help distinguish competent pilots from lucky ones and such adjunct measures can also aid in the evaluation of training interventions and the development of tactics, techniques and procedures (TTPs).

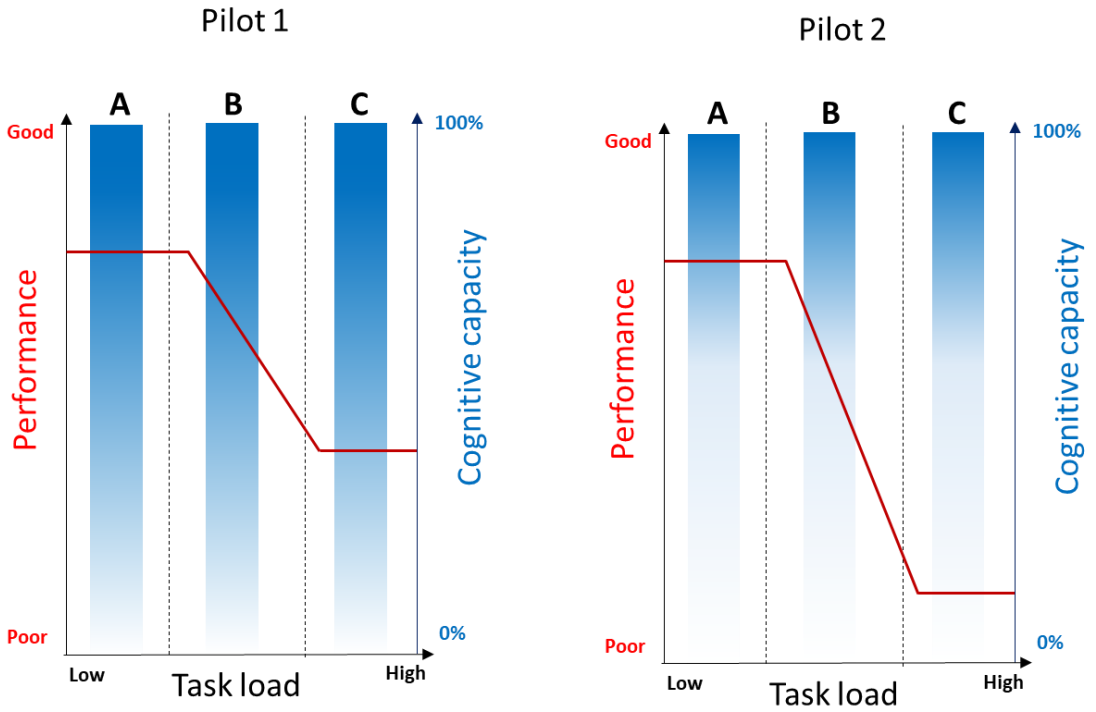

Figure 1 Hypothetical relationship between workload, performance and task load.

Aviation is also about teamwork. Even pilots of high-performance single-seat military aircraft seldom fly alone. As a minimum they operate as a pair (lead and wingman) but more usually as a four-ship (flight). On the civil flight deck pilots operate as a crew, not individuals. Consequently, although workload pertains to each individual pilot, both performance and SA need to be evaluated from the perspective of the individual *and* of the team. The performance of the flight is based upon a shared understanding of those involved (Team Situation Awareness – TSA). Furthermore, evidence now

suggests that the relationship between TSA and performance is not a simple linear relationship, but is curvilinear (Mansikka et al., 2021d; Mansikka et al., 2023).

The measurement of performance alone will only tell part of the story. For a complete picture, performance, workload, SA and/or TSA all need to be assessed, but the measurement of these factors poses practical and theoretical challenges, especially in the highly complex and dynamic environment encountered in aviation.

Measuring Situation Awareness and Team Situation Awareness

What is SA and TSA?

There are many definitions of SA, but all suggest that it is a dynamically updated mental model containing activated knowledge about a situation. It is a pilot's understanding of 'what is going on'.

Endsley's three-level model is perhaps the most dominant theory in explaining SA. Endsley (1995a) defines SA as "…*the perception of the elements in the environment within a volume of time and space* [SA level 1], *the comprehension of their meaning* [SA level 2], *and the projection of their status in the near future* [SA level 3]" (p. 36). In Endsley's model, each SA level is built upon the level below such that poor SA at a lower level contributes to low SA at higher levels (e.g. Endsley & Garland, 2000). Endsley's approach to the assessment of SA forms the basis of the developments in measurement methodology described in the following sections.

Problems in Measurement

The dominant paradigm for the assessment of SA adopts a behavioural approach, usually implemented in a flight simulator. SAGAT (Situation Awareness Global Assessment Technique) developed by Endsley (1988) uses a series of memory probes developed by subject matter experts (SMEs) that are employed during simulation scenarios. The probes are derived from an SA requirements analysis specific to the task undertaken. At various points, the simulation freezes and screens blank. Pilots are presented with a series of questions relating to Endsley's Level 1, 2 or 3 SA components. Answers are compared to the ground truth, derived from the simulation scenario to provide a measure of SA. The main disadvantage with this approach is that it frequently interrupts the simulation scenario. It has also been criticised as merely a test of memory and not of SA. This technique is also time intensive; it requires dedicated software support and the results produced are scenario specific. Such an approach may also alert pilots to the SA requirements of which they were originally unaware (Stanton et al., 2013).

Table 1 **Exemplar list of concepts and attributes for beyond visual range air combat mission** (reprinted from: Mansikka, H., Virtanen, K., Uggeldahl, U. & Harris, D. (2021). Team Situation Awareness Accuracy Measurement Technique for Simulated Air Combat - Curvilinear Relationship Between Awareness and Performance. *Applied Ergonomics, 96* (October), 103473).

Concepts	**Attributes**
Own flight/ flight members/ other friendly aircraft	Position
	Flight parameters
	Offensive capabilities
	Defensive capabilities
	Limitations
	Objectives
	Tasks
	TTPs
	Weapon effects
	Electronic warfare effects
Non-friendly aircraft	Position
	Types
	Offensive capabilities
	Defensive capabilities
	Targeted/untargeted statuses
	Declarations
	Objectives
	Tactics and manoeuvres
	Weapon effects
	Electronic warfare effects
Friendly and non-friendly forces (other than aircraft)	Positions
	Types
	Offensive capabilities
	Defensive capabilities
	Limitations
	Activity
	Electronic warfare effects
Environment	Airspace restrictions (air coordination order)

	Terrain
	Meteorological conditions (visibility, rain, etc.)

To address some of these measurement issues Mansikka et al. (2021d) developed an approach based upon Endsley's SAGAT technique and a shortened form of the Critical Decision Making (CDM) structured interview approach (Crandall et al., 2006). The approach also involves undertaking a SA requirements analysis, identifying SA concepts and attributes for a particular task. An 'attribute' is the smallest unit of task-related knowledge that a pilot can have awareness of. A 'concept' is a functional collective of attributes (Langan-Fox et al., 2004). The CDM-based interviews were undertaken during post-sortie structured interviews to derive SA scores at each level.

Initially, a list of SA concepts and attributes for beyond visual range air combat was derived from an extensive literature review (see Table 1). The content validity of this list was assured by further appraisal using operational test and evaluation pilots. This was followed by a large sample of combat ready pilots rating each component concerning how important it was to have accurate knowledge about that attribute to develop and maintain SA. To produce the final list of concepts and attributes, experienced weapons instructors and operational test pilots further reviewed the ratings. The final list was organised hierarchically such that there were seven top-level concepts, each consisting of several lower-level attributes (see Mansikka et al., 2021d).

During a post-sortie debrief, an Instructor Pilot (IP) reconstructed the mission using facilities such as cockpit video recordings, simulated flight trajectories, sensor tracks, and weapon simulations. At certain times the IP paused the review to analyse a significant decision point. The IP would introduce the decision point and the first attribute associated with it. To establish level 1 SA the pilot was asked what they understood about the attribute. Deepening probes were needed if the answer could not be determined unambiguously from the interviewee's responses. Their answers were compared against the ground truth from the simulation to establish the accuracy of their level 1 SA.

The CDM interview continued to establish the pilot's SA about the attribute's meaning with respect to the overall situation (level 2). To do this the IP drew heavily on the deepening probes developed (see Table 2) as in this case SA could not be determined from observations from the simulation, and the pilot's SA about the attribute's meaning may have been tacit. Based on the pilot's responses to these probes, the IP determined the pilot's level 2 SA. Finally, level 3 SA was established by further using the probes to help the pilot to compare and verbalise their expectations and the way in which the situation evolved. At all three levels, accurate SA, i.e., where the pilot's cognitive model of the situation corresponded to the ground truth was scored '1', whereas inaccurate SA was scored '0'. This

procedure was repeated until all relevant attributes associated with the DP were dealt with. The review was continued until the next DP was identified. Once the whole mission was reviewed and the pilots' SA of all attributes in the identified DPs were scored, SA scores were aggregated and summed to provide the overall SA index.

Table 2 **CDM deepening probes to elicit SA** (reprinted from: Mansikka, H., Virtanen, K., Uggeldahl, U. & Harris, D. (2021). Team Situation Awareness Accuracy Measurement Technique for Simulated Air Combat - Curvilinear Relationship Between Awareness and Performance. *Applied Ergonomics, 96* (October), 103473).

Probe type	**Probe content**
Information	What information were you seeking and from where? What information, if any, did you combine to gain the necessary information? How reliable was the source information? What information, if any, was missing or conflicting? What information, if any, did you misinterpret and how? Did the information change the way you understood the situation, and how?
TTPs and options	How well did the environmental cues match with TTPs? What feasible TTPs did you identify? What TTP did you select and why? / Would you have selected a different TTP than the one that was directed and why? If the TTP was directed to you, did you know what it was? What was your understanding about the flight's TTP adherence and TTP progress? What contingency TTPs, if any, were you prepared to execute and why? What were the cues that you used as triggers for a contingency TTP and why?
Goals, priorities	What were your priorities during this incident and why? What were you trying to achieve and why?
Physical/time demand	If you experienced time/physical demand, how did it affect you?
Limitations/alibies	If you experienced perceptual/technical/cognitive limitations, what were they and how did they affect you?

Expectations	Compared to your expectations, how did the status of the attribute or the situation as a whole evolve? How did the mission brief prepare you for this incident?

This approach assessed individual SA in a well-defined scenario, however Mansikka et al. (2021d; 2023) extended this approach to address Team Situation Awareness (TSA). TSA is more complex than individual SA. Endsley (1995b) argued that team members required the necessary SA for factors relevant for their specific tasks. As a result, good TSA was dependent upon team coordination and communication. almon et al. (2008) went further suggesting that TSA comprised the SA of individual members, their shared SA and the combined SA of the team (the 'common picture'). The measurement of TSA faces the same challenges as that of individual SA, with many techniques being based around approaches that require pausing the simulation to collect the data (Bolstad & Endsley, 2003; Cooke et al., 1997; Sulistyawati et al., 2009) which is undesirable and also impossible during a live exercise. Furthermore, the emphasis has generally been on determining the accuracy of TSA. However, Mansikka et al. (2023) argue that there are two components to TSA: accuracy, which assesses how closely a team's collective knowledge is aligned with the ground truth, and similarity which represents the degree of alignment of a team's collective knowledge. If TSA accuracy is high, it will closely resemble the ground truth *and* the SA of each team member will also be very similar. However, if TSA accuracy is low, the pilots may have similarly or dissimilarly inaccurate SA.

Mansikka et al. (2021d; 2023) describe the determination of TSA accuracy and TSA similarity for a flight. Each pilot's SA level 1 accuracy regarding an attribute is scored by comparing their SA with the ground truth. A higher score reflects a higher accuracy. Next, the similarity of pilots' SA is determined by making pairwise comparisons of SA between all members. Pairwise comparisons are scored such that a higher score is associated with a higher similarity. Both procedures are repeated for every attribute in an incident and for each SA-level. Level 1-3 TSA accuracy scores for an incident are determined by calculating the average of individual SA accuracy scores in respective SA levels. Level 1-3 TSA similarity scores are determined in the same fashion by calculating the average of dyads' SA level 1-3 similarity scores for an incident. Individual SA and TSA scores are calculated for every incident in the mission. Finally, an overall TSA accuracy and TSA similarity indices were determined by averaging the TSA accuracy and TSA similarity scores.

Both Mansikka et al. (2021d) and Mansikka et al. (2023) demonstrated a curvilinear (non-linear) relationship between TSA and flight performance in simulated air combat engagements. The rate of gains in offensive and defensive performance both decreased with increases in overall TSA accuracy.

The greatest performance benefits accrued with the initial increases in TSA. The relationship between a flight's TSA accuracy and performance was stronger at level 1 and weaker at level 3. This is consistent with Endsley's theory (Endsley, 1995a) as SA level 3 cannot be achieved unless the pilot has level 2 SA, which itself is predicated upon achieving level 1. With regard to TSA similarity, successful engagements were characterised by higher degrees of similarity across the flight. When a flight had gained a tactical advantage, they could control the engagement, making it easier for them to maintain a high TSA. In contrast, when a flight had lost the advantage there was a likelihood of them becoming reactive and regaining lost TSA became difficult meaning that effective decision-making suffered compromising the flight's performance.

The assessment of TSA provides insights into team performance over and above the measurement of individual SA. Many aeronautical tasks are undertaken as a team and not as an individual. However, to gain a more complete picture of the Human Factors underpinning performance, the individual cognitive load on each member of the team also needs to be assessed. Doubling the number of team members does not mean that collectively twice as much work can be undertaken as there is a processing overhead involved with communication and coordination, essential tasks to promote TSA. As a result, to understand performance the concomitant measurement of individual workload is also necessary.

Measuring Mental Workload

There is no universally accepted definition of workload, but it is generally defined as the information processing 'cost' of performing a given task, hence it is intimately related to information processing theory and the capacity of cognitive resources in working memory (Harris, 2011).

Moray (1988) suggested that there are three basic approaches to the measurement of workload: behavioural, physiological and subjective. Harris (2011) added a fourth category of 'analytical approaches', however for practical purposes, the most commonly applied approaches are subjective workload scales and physiological measures. Mansikka et al. (2019b) demonstrated a degree of convergence in workload measurement between physiological and subjective methods.

Workload measurement using subjective scales

Subjective workload measurement employs less complex uni-dimensional scales, often using a modified Cooper-Harper format (e.g. the Bedford scale: Ellis & Roscoe, 1982), and multi-dimensional measures (e.g. the Subjective Workload Assessment Technique – SWAT; Reid & Nygren, 1988: NASA Task Load Index – TLX; Hart & Staveland, 1988). Subjective measures of workload reflect the user's

experience of workload. The basic assumption is that if a pilot experiences high workload, then they are under high workload regardless of indications from other measures.

The dynamic measurement of workload shares many issues with that of (T)SA. Intrusive workload measurement using complex, multi-dimensional subjective scales that require a simulation scenario to be paused for their completion can negatively impact on primary task performance. If used concurrently with a task, multi-dimensional scales effectively present a secondary task, which has a negative impact on primary task performance. Less complex uni-dimensional scales may be more acceptable to use concurrently with a flight task, but still intrude on primary task performance to a lesser degree. However, as a result of their simplicity to complete, they have poor diagnosticity.

Although multi-dimensional scales are unacceptable to be completed in-flight (real or simulated) because of the time and effort to complete them, these types of measure have much enhanced diagnosticity which allows a more forensic analysis of the determinants of workload. These scales may be completed post-task, but then various issues arise concerning pilots' recollection of the cognitive load that they experienced during the scenario. The ratings provided may represent an assessment of the averaged workload experienced or workload peaks.

NASA-TLX (Hart & Staveland, 1988) is the most commonly used multi-dimensional scale to assess workload. The NASA TLX requires ratings to be made on three explicit dimensions relating to the sources of workload; 'mental demand' (how mentally demanding was the task); 'physical demand' (how physically demanding was the task) and 'temporal demand' (how hurried or rushed was the pace of the task) plus three further dimensions concerning the interaction of the pilot with the task: 'performance' (how successful were you in accomplishing what you were asked to do); 'effort' (how hard did you have to work to accomplish your level of performance) and 'frustration' (how insecure, discouraged, irritated, stressed, and annoyed were you)? The contribution of each of these dimensions to workload is then derived by making a series of 15 pairwise comparisons, based upon the premise that different sources of workload contribute different amounts to the overall workload in different circumstances. Participants are required to indicate which of two sources of workload is more important for the task being considered (scored '1' and '0') and the results are summed for each NASA-TLX dimension. The ratings for each sub-scale reflecting the perceived magnitude of workload for a given task are then multiplied by their associated importance weightings derived from the pairwise comparison process and summed to provide an overall workload score. This approach also enhances the diagnosticity of the instrument.

However, despite being used by numerous researchers for many years, some fundamental issues have been identified in the manner by which the weighting factors derived from the pairwise comparison

process contribute to the calculation of overall workload and support the diagnosticity of the instrument (Virtanen et al., 2022). As a result of the pairwise comparison process, it is not possible to express two (or more) workload dimensions as being equally important in their contribution to overall workload. If pairwise comparisons are conducted consistently, there exists only one possible order of the relative importance of the dimensions, with a weight of 0.33 (the maximum possible) always being allocated to the most important dimension and a weight of 0.00 to the least important dimension. As result, the pairwise comparisons, if conducted consistently, essentially ignore one of the dimensions and make the NASA-TLX a five-dimensional rating scale even if the pilot gives the dimension receiving a zero weight a workload rating. The weighting process can also lead to the inconsistent weights, with one dimension being directly considered to be more important than its contrasting dimension in the pairwise comparison process, but being deemed less important when the overall weighting order is derived.

Various enhancements to the NASA-TLX to overcome the shortcomings inherent in the original weighting system have been proposed, for example the Analytic Hierarchy Process - AHP (Saaty, 2000) and Swing method (von Winterfeldt & Edwards, 1985). Both allow weights greater than 0.33 for an individual dimension and also avoid the potential for a dimension receiving an unintentional zero weighting (see Virtanen et al., 2022). Both weighting approaches overcome the logical inconsistencies of the original weighting procedure while retaining the diagnostic benefits of the NASA -TLX method and in the case of Swing, it is also easier to administer. Moreover, in a time dependent decision environment, Swing has been found to provide stable weights over time (Lienert et al., 2016) and has shown test-retest reliability (Bottomley & Doyle, 2001). However, it does produce slightly less variance in the load dimension weights derived, compared to the AHP and the traditional NASA-TLX approach (Virtanen et al., 2022).

The NASA-TLX has also been used widely without using the weighting procedure (so-called ‘raw’ TLX) with an overall workload score being produced simply by averaging the ratings over the dimensions. This is a valid solution only if it can be assumed that the contribution to overall workload of each of the dimensions is roughly equal. If not, it will result in biased workload estimates. The ‘raw’ TLX is essentially a special case of the ‘traditional’ NASA-TLX where extremely inconsistent pairwise comparisons result in an equal weight for each dimension.

Physiological workload measurement

Variations in arousal and the general activation of the autonomic nervous system result in physiological changes which make them suitable as measures of workload. Physiological measures have the advantage of being able to provide continuous, real-time monitoring of the state of the pilot

(Jorna, 1993). Furthermore, they are passive measures that do not intrude upon performance as such, but the associated instrumentation of some measures may still be intrusive. Physiological workload measures include electrodermal activity, electroencephalography, functional near-infrared spectroscopy (fNIR), respiration rate, pupillary diameter and eyeblink (see Mansikka et al., 2016a). However, many of these measurement approaches are difficult to implement in a simulator as a result of complexities of setting up and calibrating the equipment, and other aspects of the pilot's activities during the flight scenario, for example movement, respiration/talking, and high physiological workload which result in changes in body temperature and sweat production not related to workload. These measures are almost impossible to implement in flight.

The heart is also under the control of the autonomic nervous system. Time domain-based measures of workload are derived from inter-beat intervals (IBIs). The premise is that when a pilot is under higher workload his/her heart beats slightly more quickly as higher brain activity requires a small increase in energy expenditure (these increases in blood flow in the brain can be observed directly using fNIR). Measures based upon IBI can take forms of varying sophistication, from simple measures such as mean IBI; Heart Rate (HR) and Heart Rate Variation/Variability (HRV), to more sophisticated indices of workload, for instance the square root of the mean-squared differences between successive IBIs; number of successive IBI pairs that differ by more than 50 ms, or even the integral of the IBI density distribution divided by the maximum of the distribution. IBI data can be collected relatively easily using many commercially available wearable devices, such as smart watches or sensors integrated into chest belts, which may also collect respiration data. While heart-rate data collection itself is easy, to be meaningful it must also be linked directly to the pilot's activities in the scenario (e.g. simulator logs) which can be more of a challenge.

Collecting the full ECG (electrocardiogram) waveform allows for further workload measures to be derived from the frequency domain. Frequency-based measures have proven to be sensitive to fluctuations in workload particularly in the mid-frequency band between 0.07-0.14 Hz, which is related to the short-term regulation of blood-pressure, and in the high-frequency band between 0.15-0.50 Hz, associated with respiratory functions. Decreases in power in these bands are associated with increasing workload (Mulder, 1992; Jorna, 1993; Veltman & Gaillard, 1993).

However, collecting any form of ECG data is not straightforward. There are considerable differences in individual cardiac activity and responses to varying task demand. As a result, comparisons are required both within each subject as well as across subjects (Roscoe, 1993). As within-subject comparisons are required, a resting baseline HR/HRV must also be obtained prior to any trial, ideally also followed by a post-trial resting baseline. To make the calculation of HRV meaningful,

measurement epochs of at least three minutes are required with a sampling rate in excess of 125 Hz (Lee et al., 2022). The ECG trace needs to be synchronised with events in the flight trial to corroborate the reason for any HR/HRV response observed. The trace will also need careful inspection after data are collected to remove spurious signals/artefacts (e.g. those resulting from movement or electrode motion).

In terms of the study design, differences in workload conditions need to be relatively large to show any significant differences in workload using ECG-derived measures. While being unobtrusive, cardiac-based measures are relatively insensitive. Mansikka et al. (2019b) showed that ECG-derived measures were less sensitive to differences in conditions than were subjective workload scales. Mansikka et al. (2016a, b) also found the simple measures based upon IBIs to be more sensitive to variations in workload demands than the more sophisticated ECG-derived measures described earlier. Moreover, such measures are not particularly diagnostic and can be confounded by factors unrelated to task demands. Interpretation of cardiac-based measures is often done in retrospect, limiting their utility. Reliability is often poor but can be improved by baselining measures against resting measures (or reference tasks) and by the very careful collection of data. As a result, to have any utility, physiological workload measures need to be supplemented with other performance and/or workload metrics.

Combined (T)SA, workload and performance measurement in practice

To analyse and understand performance it is essential to have an integrated underlying theoretical model describing the relationship between workload, (T)SA and performance. Mansikka et al. (2021a, b, c) developed a model of team performance and associated measurement framework to guide the collection of workload, (T)SA and performance data for the analysis of air combat engagements involving a flight of fighter aircraft. The performance aspects of the model encompass both system Output Performance – OP ('kills' and 'survival') and Normative Performance – NP (based upon adherence and execution of individual components of air combat TTPs). A simplified version of this model is described in Figure 2. While ultimately important, OP can be a poor indicator of actual performance as it is a product measure in contrast to NP, which reflects process, hence both need to be assessed. NP is more diagnostic for training purposes than OP. Mansikka et al. (2021e) describe an approach to the assessment of NP.

The air combat system model and measurement framework presented in Figure 2 describes how the selection of the TTPs to be executed is dependent upon the TSA of the flight. The accuracy of TSA is evaluated objectively from a comparison of the flight's mental model of the developing situation versus the objective situation (i.e. the ground truth) as derived from the simulation records. However,

there is a cognitive cost to developing TSA, which is reflected in the assessment of workload. Execution of the selected TTP is evaluated against the standards for NP, which also incurs a workload overhead.

Mansikka et al. (2021b, c) described how the measurement framework can be used for the testing, development and evaluation of air combat TTPs. As the approach involves live (L), virtual (V) and constructive (C) simulations, it placed differing demands on the assessment of both SA and workload. At the initial C-stage, a simulation model is used to provide estimates of the probabilities of survival (Ps) and kill (Pk) based upon the proposed TTPs but without considering the human component. At the V-stage, pilots implement the TTPs against virtual or constructive red (enemy) aircraft in simulators. Pk and Ps are calculated from results which are complemented by measures of pilots' (T)SA and workload. V-simulations provide a safe, practical and relatively inexpensive environment for the test and evaluation of the TTPs, and enable the measurement of (T)SA and workload immediately after a simulated engagement, using all the de-brief tools available. V-simulations also allow for multiple simulation runs, if required.

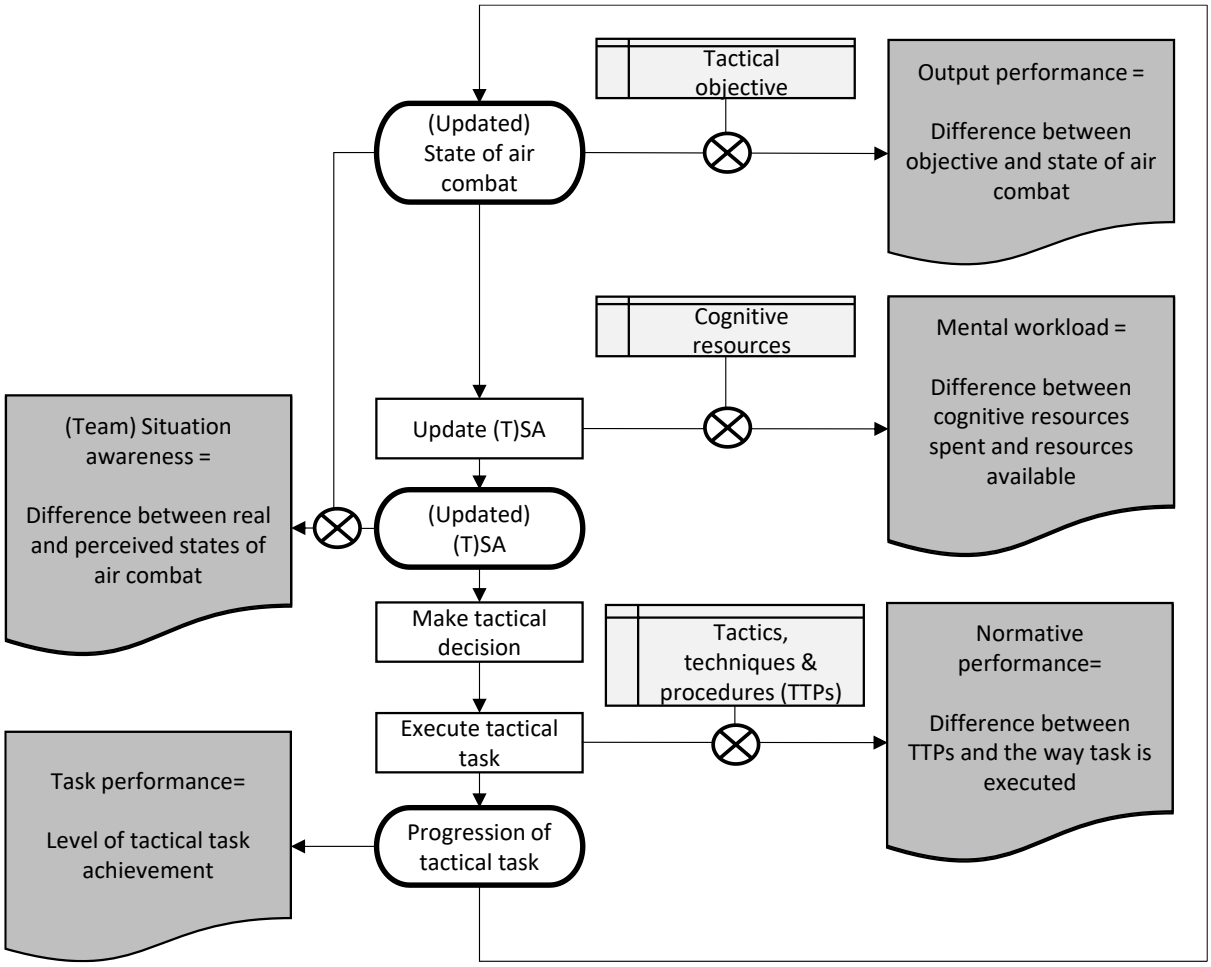

Figure 2 Simplified model of the relationship between workload, (T)SA and performance in air combat (based upon Mansikka et al., 2021a).

After further modification of the TTPs (if required) at the L-stage, pilots fly engagements in real aircraft in a real environment. L-simulations are expensive and resource heavy and are the final stage in the development process. They are important as they present pilots with real-life task complexity and stressors but are essentially used to validate the feasibility of the developed TTP in a near-real-world environment. L-simulation provides challenges for the collection of (T)SA and workload data. L-simulations cannot be paused and the time between TTP trials and data collection can be considerable. Data collection has to be unobtrusive if undertaken in real time or compromises have to be made if collected post-trial. In L-simulations, the emphasis changes away from TTP development (C- and V- stages) to verification of performance, (T)SA and level of workload imposed on the pilots.

Last words

Collecting human performance data alone is of limited utility: it provides very little explanation. Why a certain level of performance has, or has not, been achieved is often unclear. Although the relationship between performance, SA and workload can be complex, collecting all three types of data can provide a much richer description of pilot performance, providing more diagnostic data. Collecting workload and SA data is always a compromise between interfering with an ongoing task and gathering high quality information that can inform research and development efforts. However, by applying the right data gathering techniques at the right time and using a variety of complementary measures a balance between these sometimes conflicting requirements can be found.

References

Bolstad, C., & Endsley, M. (2003). Measuring Shared and Team Situation Awareness in the Army's Future Objective Force. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 47*(3), 369-373. https://doi.org/10.1177/154193120304700325

Bottomley, P., & Doyle, J. (2001). A Comparison of Three Weight Elicitation Methods: Good, Better, and Best. *Omega 29*(6), 553–560. https://doi.org/10.1016/S0305-0483(01)00044-5

Cooke, N., Stout, R., & Salas, E. (1997). Broadening the Measurement of Situation Awareness Through Cognitive Engineering Methods. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 41*(1), 215-219. https://doi.org/10.1177/107118139704100149

Crandall, B., Klein, G., Klein, G., & Hoffman, R. (2006). *Working Minds: A Practitioner's Guide to Cognitive Task Analysis*. MIT Press.

Ellis, G. A., & Roscoe, A. H. (1982). *The Airline Pilot's View of Flightdeck Workload: A Preliminary Study Using a Questionnaire. Technical Memorandum FS(B) 465*. Bedford: Royal Aircraft Establishment. https://apps.dtic.mil/sti/pdfs/ADA116314.pdf

Endsley, M. R. (1995a). Toward a Theory of Situation Awareness in Dynamic Systems. Human Factors, 37(1), 32-64. https://doi.org/10.1518/001872095779049543

Endsley, M. R. (1995b). Measurement of Situation Awareness in Dynamic Systems. Human Factors, 37(1), 65-84. https://doi.org/10.1518/001872095779049499

Endsley, M. R. (1988). Situation Awareness Global Assessment Technique (SAGAT). In Proceedings of the IEEE 1988 National Aerospace and Electronics Conference, Dayton, OH, 23–27 May 1988, (pp. 789–795). IEEE. https://doi.org/10.1109/NAECON.1988.195097

Endsley, M. R. & Garland, D. J. (2000). *Situation Awareness: Analysis and Measurement*. Lawrence Erlbaum Associates.

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, 63*(1), 124–150. https://doi.org/10.1177/0018720819875376

Fracker, M. L. (1991). Measures of Situation Awareness: An Experimental Evaluation. *Report No. AL-TR-1991-0127.* Wright-Patterson Air Force Base. Human Engineering Division, Armstrong Laboratory . https://apps.dtic.mil/sti/pdfs/ADA262732.pdf

Harris, D. (2011). *Human Performance on the Flight Deck*. Ashgate.

Hart, S. G. & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock and N. Meshkati (Eds) *Human Mental Workload* (pp. 139-183). Elsevier Science Publishers.

Jorna, P. G. A. M. (1993). Heart Rate and Workload Variations in Actual and Simulated Flight. *Ergonomics 36*(9), 1043–1054. https://doi.org/10.1080/00140139308967976

Langan-Fox, J., Anglim, J. & Wilson, J. (2004). Mental Models, Team Mental Models, and Performance: Process, Development, and Future Directions. *Human Factors and Ergonomics in Manufacturing & Service Industries 14*(4), 331–352. https://doi.org/10.1002/hfm.20004

Lee, K. F. A., Chan, E., Car, J., Gan, W.-S., & Christopoulos, G. (2022). Lowering the Sampling Rate: Heart Rate Response during Cognitive Fatigue. *Biosensors, 12*(5) 315. https://doi.org/10.3390/bios12050315

Lienert, J., Duygan, M. & Zheng, J. (2016). Preference Stability over Time with Multiple Elicitation Methods to Support Wastewater Infrastructure Decision-Making. *European Journal of Operational Research 253*(3), 746–760. https://doi.org/10.1016/j.ejor.2016.03.010

Mansikka, H., Simola, P., Virtanen, K., Harris, D., & Oksama, L. (2016a). Fighter Pilots' Heart Rate, Heart Rate Variation and Performance during Instrument Approaches. *Ergonomics, 59*(10), 1344-1352. https://doi.org/10.1080/00140139.2015.1136699

Mansikka, H., Virtanen, K. & Harris, D. (2019a). The Dissociation Between Mental Workload, Performance and Task Awareness in Fast Jet Pilots. *IEEE Transactions on Human-Machine Systems, 49*(1), 1-9. https://doi.org/10.1109/THMS.2018.2874186

Mansikka, H., Virtanen, K. & Harris, D. (2019b). Comparison of NASA-TLX scale, Modified Cooper-Harper Scale and Mean Inter-Beat Interval as Measures of Pilot Mental Workload During Simulated Flight Tasks. *Ergonomics, 62*(2), 246-254. https://doi.org/10.1080/00140139.2018.1471159

Mansikka, H., Virtanen, K. & Harris, D. (2023). Accuracy and Similarity of Team Situation Awareness in Simulated Air Combat. *Aerospace Medicine and Human Performance, 94*(6), 1-8. https://doi.org/10.3357/AMHP.6196.2023.

Mansikka, H., Virtanen, K., Harris, D., & Jalava, M. J. (2021a). Measurement of Team Performance in Air Combat – Have We Been Underperforming? *Theoretical Issues in Ergonomic Science, 22*(3), 338-359. https://doi.org/10.1080/1463922X.2020.1779382

Mansikka, H., Virtanen, K., Harris, D. & Salomäki, J. (2021b). Live-Virtual-Constructive Simulation for Testing and Evaluation of Air Combat Tactics, Techniques and Procedures, Part 1: Assessment Framework. *Journal of Defense Modeling and Simulation, 18(4),* 285-293. https://doi.org/10.1177/1548512919886375

Mansikka, H., Virtanen, K., Harris, D., & Salomäki, J. (2021c). Live-Virtual-Constructive Simulation for Testing and Evaluation of Air Combat Tactics, Techniques and Procedures, Part 2: Demonstration of Framework. *Journal of Defense Modeling and Simulation, 18*(4), 295-308. https://doi.org/10.1177/1548512919886378

Mansikka, H., Virtanen, K., Harris, D. & Simola, P. (2016b). Fighter Pilots' Heart Rate, Heart Rate Variation and Performance during an Instrument Flight Rules Proficiency Test. *Applied Ergonomics, 56*(September)*,* 213-219. http://dx.doi.org/10.1016/j.apergo.2016.04.006

Mansikka, H., Virtanen, K., Mäkinen, L. & Harris, D. (2021e). Normative Performance Measurement in Simulated Air Combat. *Aerospace Medicine and Human Performance, 92*(11), 908-912. https://doi.org/10.3357/AMHP.5914.2021

Mansikka, H., Virtanen, K., Uggeldahl, U. & Harris, D. (2021d). Team Situation Awareness Accuracy Measurement Technique for Simulated Air Combat - Curvilinear Relationship Between Awareness and Performance. *Applied Ergonomics, 96(October),* 103473. https//doi.org/10.1016/j.apergo.2021.103473

Moray, N. (1988). Mental Workload Since 1979. *International Reviews of Ergonomics, 2*, 123-150.

Mulder, L. J. M. (1992). Measurement and Analysis Methods of Heart Rate and Respiration for Use in Applied Environments. *Biological Psychology, 34*(2-3), 205-236. https://doi.org/10.1016/0301-0511(92)90016-N

Reid, G.B., & Nygren, T. E. (1988). The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload. In P. A. Hancock and N. Meshkati (Eds.), *Human Mental Workload* (pp. 185-214). Elsevier Science Publishers.

Roscoe, A. H. (1993). Heart Rate as Psychophysiological Measure for In-Flight Workload Assessment. *Ergonomics, 36*(9), 1055-1062. https://doi.org/10.1080/00140139308967977

Saaty, T. (2000). *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process*. RWS Publications.

Salmon, P., Stanton, N., Walker, G., Baber, C., Jenkins, D., McMaster, R., & Young, M. (2008). What Really Is Going On? Review of Situation Awareness Models for Individuals and Teams. *Theoretical Issues in Ergonomic Science, 9*(4), 297–323. https://doi.org/10.1080/ 14639220701561775

Stanton, N., Salmon, P. M., & Rafferty, L. A. (2013). *Human Factors Methods: A Practical Guide for Engineering and Design..* Ashgate.

Sulistyawati, K., Wickens, C. & Chui, Y. (2009). Exploring the Concept of Team Situation Awareness in a Simulated Air Combat Environment. *Journal of Cognitive Engineering and Decision Making, 3*(4), 309–330. https://doi.org/10.1518/ 155534309X12599553478791

Veltman, J. A. & Gaillard, A. W. K. (1993). Indices of Mental Workload in a Complex Task Environment. *Neuropsychobiology, 28*(1), 72-75. https://doi.org/10.1159/000119003

Vidulich, M. A. & Wickens, C. D. (1986). Causes and Dissociation Between Subjective Workload Measures and Performance. *Applied Ergonomics, 17*(4), 291-296. https://doi.org/10.1016/0003-6870(86)90132-8

Virtanen, K., Mansikka, H., Kontio, H., & Harris, D. (2022). Weight Watchers: NASA-TLX Weights Revisited. *Theoretical Issues in Ergonomic Science, 23*(6), 725-748. https://doi.org/10.1080/1463922X.2021.2000667

von Winterfeldt, D., & Edwards, W. (1985). *Decision Analysis and Behavioural Research*. Cambridge University Press.

Yeh, Y.-Y., & Wickens, C. D. (1988). The Dissociation of Subjective Measures of Mental Workload and Performance. *Human Factors, 30*(1), 111-120. https://doi.org/10.1177/001872088803000110

Biographies

Don Harris is professor of Human Factors in the Faculty of Engineering, Environment and Computing at Coventry University. He is a Fellow of the Chartered Institute of Human Factors and Ergonomics and a Chartered Psychologist. Don is the Human Factors technical advisor to Flimax ltd.

Heikki Mansikka (LtCol, ret.) is a former F/A-18 fighter pilot and an adjunct professor with the Department of Military Technology, National Defence University, Finland.

Kai Virtanen is a professor in the joint professorship of Operations Research with Systems Analysis Laboratory, Department of Mathematics and Systems Analysis, Aalto University, Finland, and the Department of Military Technology, National Defence University, Finland.