Applied Ergonomics 96 (2021) 103473

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo

Team situation awareness accuracy measurement technique for simulated air combat - Curvilinear relationship between awareness and performance

Heikki Mansikka [a,b,c,*], Kai Virtanen [a,c], Ville Uggeldahl [c], Don Harris [d]

[a] Department of Mathematics and Systems Analysis, Aalto University, P.O. Box 11100, FIN 00076, Aalto, Helsinki, Finland
[b] Insta DefSec, Tampere, P.O. Box 11100, FIN 00076, Aalto, Finland
[c] Department of Military Technology, Finnish National Defence University, Helsinki, P.O. Box 11100, FIN 00076, Aalto, Finland
[d] Faculty of Engineering, Environment and Computing, Coventry University, Priory Street, Coventry, CV1 5FB, United Kingdom

ARTICLE INFO

Keywords:
Critical decision method
Performance
Team situation awareness

ABSTRACT

A new technique for the assessment of Team Situation Awareness (TSA) accuracy based upon post task Critical Decision Method structured interviews was developed and tested using 39 combat-ready F/A-18 pilots. Pilots undertook a number of simulated air combat scenarios, flying in flights of four aircraft against a formation of enemy aircraft. Results showed a strong curvilinear relationship where high TSA accuracy resulted in higher performance in some areas of air combat, measured with friendly losses and kills. There were diminishing returns in performance as TSA accuracy increased. This may explain why previous studies on air combat have found relatively weak relationships between situation awareness and performance where the relationship has been assumed to be linear.

1. Introduction

Situation awareness (SA) has replaced traditional ‘rudder and stick’ skills as the dominant success factor in air combat (Endsley, 1995; Svenmarckt and Dekker, 2003). SA is often defined as a hierarchical three-level construct of a person’s perception of the current situation (SA level 1), comprehension of the current situation (SA level 2) and prediction of near future events (SA level 3) (Endsley, 1995).

SA as a concept can be controversial. For example, Dekker and Hollnagel (2004) have described the concept as a ‘folk model’ and adopted a reductionist approach suggesting that SA could be decomposed into measurable, specific components (e.g., decision-making, perception, understanding, and long-term memory). They also argued that it was immune to falsification (see also Flach, 1995). Even if it is accepted that SA actually exists, the scientific nature of the concept is open to debate. For example, does it reside within the cognition of the user or is it an emergent property of the wider system, and what is the most appropriate approach to its measurement (for more details, see the extensive reviews by Salmon et al., 2008; Endsley, 2015; Stanton et al., 2017; Nguyen et al., 2019)? Nevertheless, it is evident that the concept of SA has become an important metric in the evaluation of systems and human performance. As Wickens (2008) noted “… *one can speak to the increased use of the construct in both theory and applications as testimony to its viability, as well as note that such strong criticism is also an index of the value of the SA concept to human factors science*” (p. 401).

Fighter aircraft are typically operated as a flight, which is a standard fighting unit comprised of two sections, each with two pilots. While every pilot within a flight has his/her own SA about the air combat situation, the flight has also a collective Team SA (TSA). While high SA is related to good performance in this context, it cannot simply be inferred that success in combat is a direct result of high SA. The relationship between performance, workload and SA is a complex one. The relationship may either be relatively weak (e.g., Fracker, 1991; Endsley et al., 2000; Strybel et al., 2008; Endsley, 2019) or complex and unclear (Durso et al., 1998: Sulistyawati et al., 2009; Mansikka et al., 2019a; Joffe and Wiggins, 2020). When assessing performance in such a context, both outcome and process measures should be considered (Mansikka et al., 2019b, 2019c, 2020). In a highly dynamic, uncertain environment such as air combat, success may occasionally be a product of chance factors and vice versa. As a result, there is a practical need to estimate reliably a flight’s TSA in air combat, as understanding the TSA can help in separating competent teams with high TSA from lucky ones with low TSA. An estimate of TSA has also an impact on evaluating training interventions and as well as on the evaluation and comparison of the utility of tactical operating procedures, the competence of teams and/or the applicability of aircraft systems.

[*] Corresponding author. Department of Mathematics and Systems Analysis, Aalto University, P.O. Box 11100, FIN 00076, Aalto, Helsinki, Finland.
E-mail address: heikki.mansikka@aalto.fi (H. Mansikka).

https://doi.org/10.1016/j.apergo.2021.103473
Received 18 January 2021; Received in revised form 12 May 2021; Accepted 17 May 2021
Available online 2 June 2021

TSA is more complex than individual SA. Endsley (1995) argued that for the flight to meet its goals, the flight members required the necessary SA for those factors relevant for their specific duties. The major challenge to achieve good TSA is the coordination of team members. Salmon et al. (2008) went further, suggesting that TSA had other aspects to it, including SA of individual team members, their shared SA and what they described as the 'common picture', the combined SA of the whole team. In this case, the focus was upon the measurement of the shared aspect of TSA, what Harris (2011) described as 'overlapping' SA; those common SA elements a team must share for effective performance.

1.1. Assessing TSA

TSA has been assessed in several dynamic, high-risk environments using a number of approaches. Early approaches (e.g., Wellens, 1993) inferred TSA from performance in a civilian command and control task, although measures of SA were later supplemented by post-trial review structured interviews using specific memory probes. It was noted that high SA was not necessarily related to high performance. It was observed that if it took too long to develop SA, this could be at the cost of not undertaking other actions associated with good performance, hence the cost of developing high TSA outweighed its performance benefits. Fowlkes et al. (1994) in a study of US helicopter crews, also used an approach to assess TSA based upon observed behaviors. Amongst other teamworking categories, pre-determined team behaviors indicative of TSA were defined *a priori* to specific task events. Subject matter expert (SME) observers then scored team performance on the basis of the presence or absence of these behaviors in a teamwork scenario. Both these approaches had the advantages of not intruding on task performance, but both also only inferred the degree of TSA achieved from observed behaviors.

Gorman et al. (2006) described the theoretical development and initial application of a TSA measure (Coordinated Awareness of Situation by Teams - CAST) based upon the retrospective analysis of a number of military incidents where SA was 'lost'. Its proactive application in a simulated unmanned air vehicle reconnaissance experiment involving a distributed team is described in Gorman et al. (2005). This approach also used an observer-based assessment paradigm, however, it was somewhat less structured and prescriptive, as it was argued that complex, dynamic situations do not present pre-defined events to assess. It was, however, predicated upon the assumption that TSA was mostly challenged when teams faced 'unlikely events requiring adaptive and timely team-level solutions' or 'roadblocks'. TSA was evaluated when such 'roadblocks' were encountered (e.g., the failure of a communication system). The measurement process did not intrude on team performance but was not applicable in 'normal' operations.

Endsley (1998) has adopted a different method based upon the SAGAT (Situation Awareness Global Assessment Technique) approach used to assess individual SA. This approach is only applicable to simulations (see, e.g., Falkland and Wiggins, 2019; Krampell et al., 2020) as it requires the researcher to freeze the scenario at pre-determined points to probe individual team members with structured queries concerning the current and future states of the system in order to assess their overlapping SA (Bolstad and Endsley, 2003). TSA is assessed from the accuracy of team responses to SA probes, with more team members responding correctly being indicative of higher TSA.

However, there are challenges in the measurement of TSA in a dynamic air combat context. Techniques based around approaches such as that proposed by Cooke et al. (1997), Bolstad and Endsley (2003) and Sulistyawati et al. (2009) require pausing the activity for a time to collect the data, something which is not always possible, especially during a live exercise. Furthermore, Salmon et al. (2009) have criticized the validity of this approach suggesting that it is unclear if it is SA or recall memory being assessed. In addition to measures of SA, Sulistyawati et al. (2009) also used combat performance measures to assess TSA effectiveness, however, as noted earlier, this can be a misleading measure as there is often a dissociation between SA and performance (Mansikka et al., 2019a).

Self- and peer-appraisal techniques (see, e.g., Weigl et al., 2020) on the other hand, are not intrusive, but may reflect confidence rather than TSA (Lichacz, 2006) or knowledge (Prince et al., 2007). Fowlkes et al. (1994), Salas et al. (1995), Bolstad and Endsley (2003) and Gorman et al. (2006) all argue that TSA is a product of teamwork. Therefore, they essentially claim that an assessment of teamwork behavior is the best approach for the evaluation of the process of gaining TSA, rather than for its evaluation as a product or an emergent state. Rosenman et al. (2018) utilised an approach to the assessment of TSA based upon post-task probes of 3 level SA. SA was based upon the response accuracy to these questions: the TSA metric was determined by averaging the pairwise agreement for each dyad in the team.

1.2. SA as knowledge

A fighter pilots' knowledge is stored in, and transferred between, long term memory (LTM) and working memory (WM) (Atkinson and Shiffrin, 1968; Wickens, 1991). Knowledge in LTM is organised in the form of mental models (MMs), which are a collective name for the structure and content of the pilots' knowledge regarding the air combat environment and the sequence of activities of their tasks (Gilbert, 2011; Wilson and Rutherford, 1989). In this paper, an attribute refers to the smallest unit within the air combat environment that the pilots can have knowledge of, whereas a concept is a functional collective of attributes (Langan-Fox et al., 2000). For example, a non-friendly aircraft is a concept with position and type as its attributes.

Knowledge in WM is based on activated MMs which pilots use to reason and to comprehend the air combat environment (Johnson-Laird, 1983). When the activated MMs are updated with observations, the resulting dynamic knowledge is often referred to as SA (Endsley, 1995, 2000, 2019). SA is fundamentally the pilot's knowledge of what s/he thinks is happening now and in the near future, as opposed to what is actually happening and what will be really happening (Endsley, 1993; Wickens et al., 2004), i.e., the 'ground truth'.

1.3. TSA as collective knowledge

The collective knowledge possessed by a flight has been interchangeably referred to as shared knowledge (Hecker, 2012), team mental model (Klimoski and Mohammed, 1994; van der Haar et al., 2015), shared cognition (Cooke et al., 2000), TSA (Cooke et al., 2001), team situation model (Cooke et al., 2017) and shared understanding (Johnson & O'Connor, 2008). TSA is the flight members' collective SA regarding the attributes relevant for a flight.

There are numerous definitions for TSA in the literature. Endsley (1995) defined TSA as 'the degree to which every team member possesses the SA required for his or her responsibilities' (Endsley, 1995; p. 39). Salas et al. (1995) suggested that 'TSA is at least in part the shared understanding of a situation among team members at one point in time'. Wellens (1993) defined it as 'the sharing of a common perspective between two or more individuals regarding current environmental events, their meaning and projected future status'. However, all definitions encompass shared SA (Harris, 2011). Shared SA is underpinned by elements of a common mental model of the flight combined with an appreciation of the flight members' individual responsibilities. TSA builds upon the concept of Transactive Memory (Wegner, 1985) which proposes that a memory system comprised of a group of people is more complex and effective than its individual constituents. Salas et al. (2005) argued that TSA was based upon the information exchange required for the mission which determined individual tasks and roles. Therefore, TSA was the result of the interactions of individual SA (q.v. Rogers, 1997). As a result of the complex interaction between individuals' SA and the processes by which SA is shared, several authors have suggested that TSA cannot be adequately described using a reductionist perspective as it is more than the sum of its component parts (Fiore et al., 2012; Stanton et al., 2017). Gorman et al. (2006) concluded that 'TSA is beyond the scope of adding up operators' private awareness of the situation and is predicated on operator interaction' (p. 1320). This view is entirely consistent with the macrocognitive perspective adopted in the study of transactive memory and distributed cognition – both approaches containing more and better information that any one person could access. Rogers (1997) developed this further describing the generic properties of cognition in people working as a team. She suggested that a team has properties over and above those of the individuals making up a team. For example, the knowledge possessed by team members is highly variable and redundant, and distribution and access to information promotes coordinated action. Teams engage in interactions allowing them to pool their cognitive resources, they share knowledge through both formal and implicit communication, and from prior knowledge of each other. One agent may compensate for degradation in SA in another (Stanton et al., 2006). As a result, when adopting such a macrocognitive approach the emphasis in analysis changes away from the individual to that of the collective properties and performance of the team. A more holistic perspective is required. Endsley and Jones (2001) developed a model of TSA comprising four necessary components to support it. The model comprised 1) Requirements (the information and goals that need to be shared); 2) Devices (methods to share this information); 3) Mechanisms (the devices to aid in developing TSA) and 4) Processes (the formal processes for sharing information, verifying understanding, prioritising tasks and establishing contingencies).

The flight uses TSA to describe, explain and predict the progress of the air combat, and to recognise and select appropriate tactics, techniques and procedures (TTPs) (Rouse and Morris, 1986). TTPs are rules which, when followed, create discipline and provide structure to an otherwise unpredictable and chaotic activity (Rajabally et al., 2009). Accurate TSA is an essential contributor to the performance of a flight (Converse et al., 1991; Langan-Fox et al., 2004). TSA accuracy refers to the level of agreement between the flight members' SA regarding the attributes and the objective reality (the ground truth).

When TSA is considered as collective knowledge, estimation of TSA essentially becomes a knowledge elicitation task. Knowledge elicitation is a 'component of knowledge acquisition in which information pertaining to the reasoning and other thought processes needed to perform a job is obtained from a human source' (Fowlkes et al., 1994). There exists a strong research tradition in the field of knowledge elicitation, which provides alternative approaches to estimate collective knowledge, or TSA, in an air combat context (see, e.g., Cooke et al., 2000; Hoffman et al., 1998; Hoffman et al., 1995; Cooke, 1994; Klein et al., 1989).

1.4. *Aims and objectives*

The aims of this paper were twofold: firstly, to describe the relationship between the accuracy of TSA and performance of a flight in air combat missions of increasing levels of cognitive demand required to develop and maintain TSA, and secondly to overcome the requirement to interrupt the task during the time when data are collected. Once an estimate of the flight's TSA accuracy in a natural air combat environment had been obtained, the objective of this paper was to provide a novel explanation for the frequently observed dissociation between performance, measured using friendly losses and kills, and TSA.

The remainder of the paper is organised as follows. Section 2 describes how the TSA accuracy measurement technique was developed. Section 2.1 illustrates how the TSA concepts and attributes were identified and Section 2.2 describes how the pilots' knowledge is elicited. Section 2.3 explains how TSA accuracy is estimated. Sections 3 demonstrates the use of the TSA accuracy measurement technique in a real-life setting, followed by the associated results in Section 4 and discussion in Section 5.

2. TSA accuracy measurement technique

2.1. *Attribute development*

Estimating the flight's TSA accuracy in a beyond visual range (BVR) air combat mission is essentially about determining their knowledge about the relevant attributes within this air combat environment. To develop a TSA accuracy measurement technique that could be used across different live and virtual BVR missions, it was necessary to ascertain which concepts and attributes were relevant in a typical BVR mission. This was carried out using a phased technique based upon that used by Langan-Fox et al. (2000, 2004) and Johnson et al. (2007).

First, some 200 research articles (e.g., North Atlantic Treaty Organization (NATO), 1995; Endsley and Garland, 1995; Gilson et al., 1994; Endsley, 1993), technical reports (e.g., Vidulich et al., 1994; Fracker, 1991) and air force manuals (e.g., Korean Air Force, 2005; Royal Norwegian Air Force, 2001) were reviewed. Relevant material was searched using Google Scholar (https://scholar.google.com), and the search facilities of the NATO Science and Technology Organization (https://www.sto.nato.int/Pages/default.aspx) and Defence Technical Information Center (https://discover.dtic.mil). Based on this review and an analysis of the flight's BVR task, an initial list of 298 relevant concepts and attributes was formed. The initial list of concepts and attributes was reviewed by experienced instructor pilots (IPs) who shortlisted a list of concepts and attributes by removing duplicates and combining the remaining items into meaningful units. The content validity of the resulting potential list was evaluated by operational test and evaluation (OT&E) pilots. Next, 61 combat ready F/A-18 pilots (mean age 32.6 years, SD = 3.7) were recruited from two fighter squadrons to rate the attributes. All pilots were volunteers and were not paid for the rating task. The pilots rated each attribute in the potential list based on how important it was for the flight members to have an accurate knowledge about the attribute. Ratings were conducted with respect to a typical BVR mission. Ratings ranged from '1' (low importance) to '7' (high importance). To make the rating scale more meaningful for the pilots, the ratings were given verbal descriptions and structured into a hierarchical rating aid based around the format of the Cooper-Harper rating scale, shown in Fig. 1 (Cooper and Harper, 1969).

Eleven experienced weapons instructors and operational test pilots reviewed the ratings and produced a final list of concepts and attributes. The final list was organised hierarchically such that there were seven top-level concepts, each consisting of several further attributes. Table 1 summarises the final list. The concepts and attributes were selected and formulated such that they were platform independent.

2.2. *Post-sortie interview method*

An interviewer determines TSA accuracy during structured interviews in the post sortie debrief. During the debrief, an IP reconstructs the mission using facilities such as cockpit video recordings, aircraft's simulated flight trajectories, sensor tracks, and the weapon simulations of all participating aircraft. Due to its comprehensiveness, the reconstruction is often referred to as the 'ground truth' of the mission (Waag and Houck, 1994).

The ground truth is reviewed during the debrief. As the IP identifies a critical incident from the ground truth, the review is paused. A critical incident is defined as one which has direct flight safety implications, or which contributes to changes in the level of completion of the flight's taskwork. A flight's TTP selection decision and TTP execution are typical examples of taskwork related incidents. For a more detailed discussion about a flight's taskwork, see Joint Chief of Staff (2013) and Mansikka et al. (2020).

At this point, the execution of the debrief deviates slightly from its standard flow to accommodate the TSA elicitation probes. This aspect of the debrief is based upon a modified, shortened form of the Critical Decision Making (CDM) structured interview approach (see, Crandall

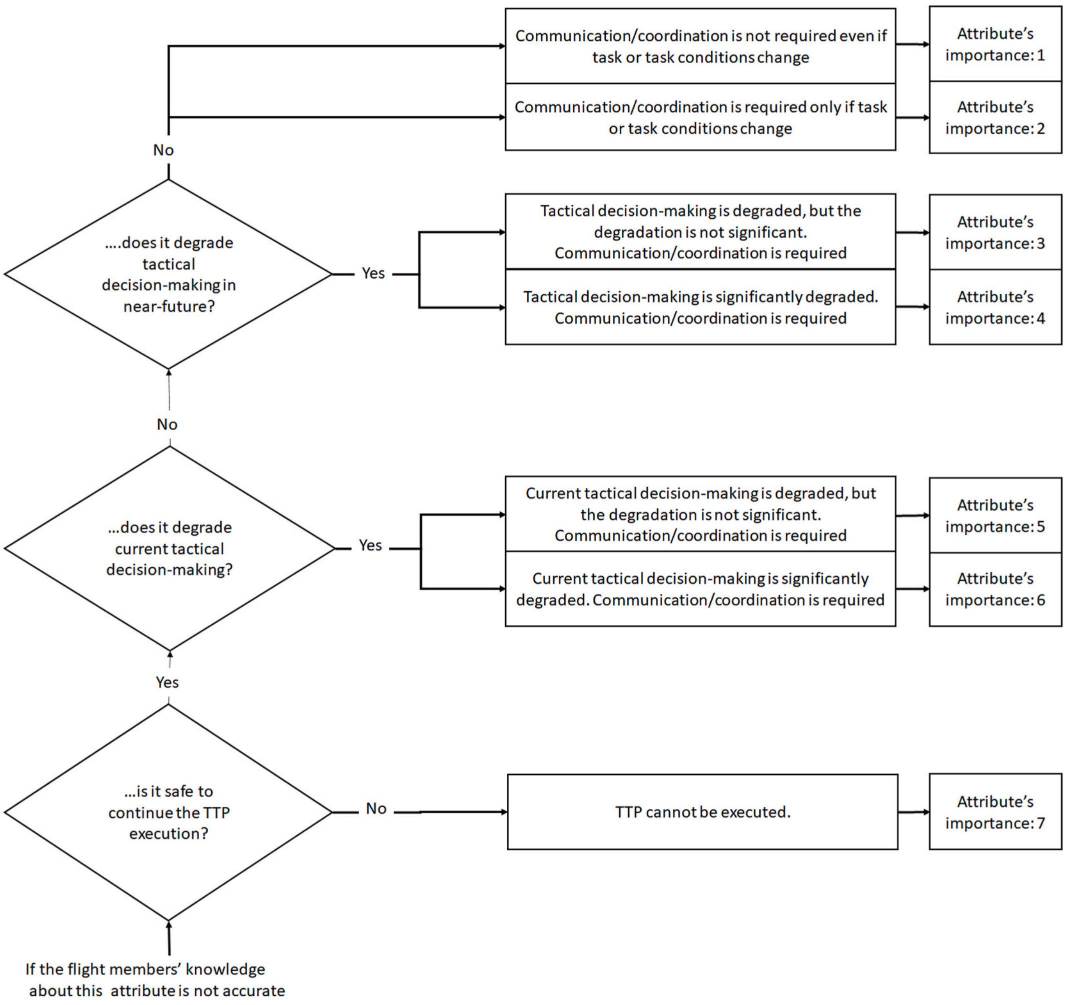
Communication/coordination is not required even if task or task conditions change
Attribute's importance: 1
Communication/coordination is required only if task or task conditions change
Attribute's importance: 2
No
....does it degrade tactical decision-making in near-future?
Yes
Tactical decision-making is degraded, but the degradation is not significant. Communication/coordination is required
Attribute's importance: 3
Tactical decision-making is significantly degraded. Communication/coordination is required
Attribute's importance: 4
No
...does it degrade current tactical decision-making?
Yes
Current tactical decision-making is degraded, but the degradation is not significant. Communication/coordination is required
Attribute's importance: 5
Current tactical decision-making is significantly degraded. Communication/coordination is required
Attribute's importance: 6
Yes
...is it safe to continue the TTP execution?
No
TTP cannot be executed.
Attribute's importance: 7
If the flight members' knowledge about this attribute is not accurate

Fig. 1. Attribute rating aid.

et al., 2006). The CDM interview is essentially a retrospective, semi-structured knowledge elicitation technique, which uses cognitive probes to extract experts' SA. This is a similar approach to that used by Plant and Stanton (2015).

The CDM interview to derive the data for the TSA accuracy assessment has four phases: incident identification, timeline construction, deepening probes (to tap tacit knowledge) and 'What if' queries. Klein and Armstrong (2004) recommend tailoring the deepening probes to meet the specific research objectives. Therefore, novel probes were developed to support TSA elicitation in a natural air combat training environment. Table 2 describes these probes.

IPs are trained to facilitate the CDM interview. All IPs taking part in the study need to be familiar with identifying critical incidents and open-ended questioning. Furthermore, they all must be briefed about the phases and objectives of the CDM interview, and the use of the deepening SA probes.

At the first step, the IP introduces the incident and the first attribute associated to it (see Table 1). The pilot is asked if s/he has correctly perceived the attribute, i.e., if the pilot has SA level 1 about the attribute. Sometimes the correctness or incorrectness of the perception can be determined directly from the ground truth and there is no need for elaboration or deepening probes. For example, if information needed to form SA level 1 about the attribute is included in a radio call and the pilot provides a correct, positive acknowledgement to that radio call, it is clear that the pilot has SA level 1 for that attribute. However, deepening probes are needed if the correct (or incorrect) perception cannot be unambiguously determined from the pilot's behavior.

Next, the pilot's SA about the attribute's meaning with respect to the situation at hand, i.e., SA level 2, is determined. To assist pilots in verbalising this aspect of SA, the IP draws heavily on the probes listed in Table 2. This is because SA about the attribute is not necessary realized as an observable behavior and the pilot's SA about the attribute's meaning may be tacit. Moreover, an observable behavior suggesting sufficient SA level 2 may be founded on insufficient or wrong SA about the ground truth. Based on the pilot's responses to the deepening probes, the IP determines the pilot's SA level 2.

Finally, the pilot's SA level 3 is established by using the deepening probes which motivate the pilot to compare and verbalise his/her expectations and the way the situation had evolved. It is not recommended to probe about the future status of the situation as by the time the probes are introduced, the pilot has not yet seen the ground truth of that situation. The only deepening probes which are explicitly forward looking, are the ones associated with contingencies. They are also the probes which are used to complete the fourth, 'What if', phase of the interview.

Once the pilots' SA levels 1–3 concerning the first attribute has been elicited, the procedure is repeated for all other attributes related to the same critical incident. Reviewing of the ground truth continues until the IP identifies the next incident, the debrief is paused and the CDM interview is re-initiated. This procedure is repeated until the pilot's SA with respect to every critical incident contained within the ground truth has been elicited. The number of attributes to be probed and the number of probes to be used to elicit each attribute depend on the time available for the debrief and the desired comprehensiveness of the knowledge elicitation.

Table 1
Final list of concepts and attributes.

Concepts	Attributes
Own flight/ flight members/ other friendly aircraft	Position Flight parameters Offensive capabilities Defensive capabilities Limitations Objectives Tasks (kill and live chains) 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)	Position Types Offensive capabilities Defensive capabilities Limitations Activity Electronic warfare effects
Environment	Airspace restrictions (air coordination order) Terrain Meteorological conditions (visibility, rain, etc.)

Table 2
Deepening probes.

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?

2.3. *Determining TSA accuracy*

An estimation of the flight's TSA accuracy for SA levels 1–3 utilise the pilots' SA levels 1–3 determined during the post-sortie interview. After the pilots' SA levels 1–3 with respect to an attribute have been obtained during the debrief, TSA accuracy scores for SA levels 1–3 are defined for that attribute. First, each pilot's SA level 1–3 accuracy for the attribute is scored. Separate scores are determined for each SA level. Accurate SA level about an attribute is scored as '1', whereas inaccurate SA is scored as '0'. Accurate SA means that a well-informed decision can be made based on that SA. The level of accuracy required to make a well-informed decision is dependent on the tactical situation. By summing the pilots' SA level accuracy scores for an attribute, TSA level 1–3 accuracy scores are obtained for that attribute. These scores can range from '0' to '4' for the flight. The procedure is repeated for every attribute. Once the TSA level 1–3 accuracy scores have been obtained for every attribute, TSA level 1–3 accuracy indices are calculated by averaging the TSA level 1–3 accuracy scores of the respective TSA levels. Finally, the TSA accuracy index of the flight is determined by averaging all TSA accuracy scores - which is essentially the same as averaging the TSA level 1–3 accuracy indices.

3. Demonstration

3.1. *Participants*

Thirty-nine qualified F/A-18 fighter pilots participated in the simulated air combat demonstration. All participants were male. The participants' average flight experience with F/A-18 aircraft was 543 flight hours (SD = 302) and they were all fit to fly. Written informed consent was obtained from each participant. A fighter controller was assigned for each flight, whose task was to support the flight according to the fighter controllers' standard operating procedures. The fighter controller's impact on flight's TSA was not assessed.

3.2. *Apparatus*

For each mission, four flight training devices (FTDs) used for air combat training in the fighter squadrons were used. Two types of FTDs were used; one had a 216° field of view and a fully functional cockpit, whereas the other had limited cockpit functionality and a virtual reality headset with a 360° field of view. The fighter controller had access to a simulated ground-based radar picture of the operating area. The fighter controller's workstation and the participants' FTDs were networked, and all participants were able to communicate via radio and datalink. All enemy, or red, aircraft were computer generated simulation entities. They were programmed with 'perfect SA' and scripted to mimic typical red tactics used in western air combat exercises.

3.3. *Procedure*

3.3.1. *Flying mission*

Participants were assigned into flights based on their training rosters. Once the flights were given their task, they conducted a standard mission brief and entered the FTDs. Each simulation consisted of a flight and red aircraft. The flight's task was to intercept the red aircraft and the red aircraft were programmed to intercept the flight. Before the simulation was started, the flight and the red aircraft were initialised at their designated starting positions, speeds and altitudes. The red presentation in each mission placed equal cognitive demand on pilots. However, the cognitive demand required to develop and maintain TSA was manipulated between missions by varying the functionality of the datalink which is the fundamental means for building awareness of the tactical picture. With reference to Endsley and Jones's (2001) model of the mechanisms underpinning TSA, the datalink is a device which aids in building TSA and holds a partial representation of the tactical picture. The greater the functionality of the datalink was restricted, the more the pilots had to rely on radio communications and a shared appreciation of mission objectives and TTPs to build TSA.

Three levels of cognitive demand required to develop and maintain TSA were used: high, medium, and low. In the low cognitive demand condition, the datalink was fully functional and enabled digital transfer of information between aircraft, and between aircraft and the fighter controller. In the medium demand condition, the datalink allowed information transfer only between aircraft. In the high demand condition, the datalink was disabled. In all conditions, the necessary information was available but just needed different mechanisms to share it to achieve TSA. Achieving a high TSA was possible, but cognitively demanding, even in the high cognitive demand condition where the datalink was disabled, and information normally transferred via datalink had to be transferred via radio. Two different missions were prepared for each cognitive demand condition. During each simulator session, the flights flew missions with all three levels of cognitive demand. Eleven flights flew each cognitive demand condition twice, i.e., six different missions. Due to time constraints, four flights flew each condition just once, i.e., three different missions. The order in which the different cognitive demand conditions were introduced was randomised between the flights. Once the simulation was initiated, the mission was left to evolve freely until all red aircraft were destroyed, the flight was destroyed, or 10 min had elapsed.

3.3.2. *Data collection*

Once the simulation had ended, the mission was reconstructed and the debrief was initiated. The IP facilitated the debrief, determined the pilots' SA levels and defined the TSA accuracy scores. In the debrief, the ground truth was reviewed, and it was paused each time the IP identified a critical incident – typically a decision point related to TTP selection. Once paused, the pilots' SA levels 1–3 regarding the attributes associated with that incident were elicited. To assist the pilots' recall, they were let to view, replay, zoom and rewind the ground truth at their will. Whenever the pilots' SA was not clearly observable from their behavior, the IP used the deepening probes listed in Table 2 to determine their SA. Determination of SA levels 2 and 3 was particularly dependent on the use of the deepening probes. Once the pilots' SA levels 1–3 for an attribute were obtained, TSA accuracy scores for SA levels 1–3 were defined by comparing the pilots' SA about that attribute to the ground truth. The procedure was repeated for every critical incident and every attribute associated with those incidents. TSA level 1–3 accuracy indices and a TSA accuracy index were calculated from the TSA accuracy scores.

The debriefs lasted less than 60 min. On average, the IP identified 73 attributes during each debrief.

4. Results

The relationships between performance, measured with friendly losses and kills, TSA accuracy and cognitive demand were analysed. The unit of analysis was the flight (N = 15), not each individual pilot.

4.1. *Performance and TSA accuracy indices versus cognitive demand*

There was a significant difference observed in TSA accuracy indices with respect to the level of cognitive demand imposed by the mission ($F_{2,13} = 473.562$; $p < 0.001$; $\eta_p^2 = 0.971$), see Table 3. All pairwise comparisons between the cognitive demand levels and the TSA accuracy index were significant ($p < 0.001$). In general, the TSA accuracy index declined as cognitive demand increased.

There was also a significant difference observed in the number of friendly losses with respect to the level of cognitive demand ($F_{2,13} = 6.168$; $p < 0.026$; $\eta_p^2 = 0.306$), see Table 3. However, the only significant pairwise comparison was between losses in the low and high cognitive demand missions. The difference in the number of kills with respect to cognitive demand was verging on significance ($F_{2,13} = 4.249$; $p < 0.058$; $\eta_p^2 = 0.233$). No pairwise comparisons were significant.

Table 3
Means (M) and standard deviations (SD) of the TSA accuracy index, friendly losses and kills, broken down by the level of cognitive demand, N = 15.

Cognitive demand	TSA accuracy index		Losses		Kills	
	M	SD	M	SD	M	SD
Low	3.54	0.22	0.40	0.69	6.40	0.60
Medium	2.58	0.45	0.70	0.84	5.53	1.82
High	1.72	0.44	0.97	1.16	5.57	1.59

4.2. *TSA accuracy index versus performance*

At all levels of cognitive demand, there was a highly significant negative curvilinear relationship between the TSA accuracy index and friendly losses (Low cognitive demand: $R = 0.929$; $R^2 = 0.863$; $R^2_{adj} = 0.841$; $F_{2,12} = 37.955$; $p < 0.001$; Medium cognitive demand: $R = 0.821$ $R^2 = 0.674$; $R^2_{adj} = 0.620$; $F_{2,12} = 12.428$; $p < 0.001$; High cognitive demand: $R = 0.724$; $R^2 = 0.476$; $R^2_{adj} = 0.726$; $F_{2,12} = 7.353$; $p < 0.01$).

With friendly losses dependent upon the TSA accuracy index, the best fit models were all quadratic in nature (Low cognitive demand: Losses = TSA accuracy index*-56.899 + TSA accuracy index^2 * 7.734 + 104.519; Medium cognitive demand: Losses = TSA accuracy index*-5.678 + TSA accuracy index^2 * 0.849 + 9.549; High cognitive demand: Losses = TSA accuracy index*-7.581 + TSA accuracy index^2 * 1.581 + 9.023).

A higher TSA accuracy index was associated with fewer friendly losses at all levels of cognitive demand. However, the greatest benefits from TSA accuracy improvement accrued at lower levels of TSA accuracy and the return diminished as TSA accuracy improved.

At the low level of cognitive demand, there was a significant curvilinear relationship between the TSA accuracy index and kills (Low cognitive demand: $R = 0.659$; $R^2 = 0.435$; $R^2_{adj} = 0.341$; $F_{2,12} = 4.613$; $p < 0.05$). With kills dependent upon the TSA accuracy index, the best fit model was again quadratic in nature (Low cognitive demand: Kills = TSA accuracy index* 57.558 + TSA accuracy index^2 * −8.287 – 93.472). There were no significant relationships between kills and the TSA accuracy index in either the medium or high cognitive demand missions. Similar to the results for losses, the results with respect to kills in the low cognitive demand condition again showed a pattern of diminishing returns, with performance benefits reducing with increasing TSA accuracy.

The significant regression results were further decomposed to investigate the relationship of TSA level 1–3 accuracy indices with performance. A series of stepwise multiple regressions (with p to enter set at <0.05 and p to exclude > 0.10) were performed to predict performance from TSA levels 1, 2 and 3 accuracy indices.

4.3. *Low cognitive demand – Losses*

Table 4summarises the mean friendly losses and TSA level 1,2 and 3 accuracy indices at low cognitive demand. All predictor variables, i.e., TSA accuracy indices of TSA levels 1, 2 and 3, correlated highly with the number of friendly losses. All TSA predictor variables were also highly inter-correlated (Table 5).

In the low cognitive demand mission, only one predictor variable entered into the final regression equation, the TSA level 1 accuracy index, as a result of the high inter-correlations between the predictor variables. The resulting equation was, however, highly significant. There was a significant negative relationship between the friendly losses and TSA accuracy at level 1 ($R = 0.830$; $R^2 = 0.689$; $R^2_{adj} = 0.604$; $F_{3,11}$-

Table 4
Means (M) and standard deviations (SD) of friendly losses and TSA level 1, 2 and 3 accuracy indices at low cognitive demand, N = 15.

	M	SD
Losses	0.40	0.69
TSA level 1	3.67	0.21
TSA level 2	3.63	0.23
TSA level 3	3.33	0.29

Table 5
Correlations between friendly losses and TSA level 1, 2 and 3 accuracy indices (predictors) and inter-correlations between predictor variables.

	Friendly losses (low cognitive demand)	TSA level 1	TSA level 2
TSA level 1	−0.79**		
TSA level 2	−0.73**	0.88**	
TSA level 3	−0.66**	0.59*	0.81*

N = 15; *p < 0.01; **p < 0.001.

$= 8.113$; $p < 0.01$). With friendly losses dependent upon TSA accuracy index, the equation for the best fit line was: Losses = TSA level 1 accuracy index*-2.780 + 10.199. A similar pattern of results was observed in the medium and high cognitive demand missions.

4.4. *Medium cognitive demand - Losses*

Table 6summarises the mean friendly losses and TSA level 1, 2 and 3 accuracy indices at medium cognitive demand. Again, all predictor variables, i.e., accuracy indices of TSA levels 1, 2 and 3, correlated highly with the number of friendly losses (see Table 6). All predictor variables were also highly inter-correlated (Table 7).

As before, in the medium cognitive demand mission, the only predictor that entered into the final regression equation was the TSA level 1 accuracy index, but the resulting equation was highly significant. There was a significant negative relationship between friendly losses and the TSA level 1 accuracy index ($R = 0.815$; $R^2 = 0.664$; $R^2_{adj} = 0.572$; $F_{3,11} = 7.235$; $p < 0.01$). With friendly losses dependent upon the overall TSA accuracy index, the best fit line was: Losses = TSA level 1 accuracy index*-1.181 + 5.008.

4.5. *High cognitive demand – Losses*

With the exception of the TSA level 3 accuracy index in the high cognitive demand mission the two remaining predictor variables, i.e., the accuracy indices of TSA levels 1 and 2, correlated highly with the number of friendly losses (see Table 8). The remaining two TSA indices were highly inter-correlated (Table 9).

In the high cognitive demand mission, once again the only significant predictor that entered into the regression equation was the TSA level 1 accuracy index, but the resulting equation was significant. As before, there was a significant negative relationship between friendly losses and the TSA level 1 accuracy index ($R = 0.726$; $R^2 = 0.526$; $R^2_{adj} = 0.448$; $F_{3,11} = 6.670$; $p < 0.01$). With friendly losses dependent upon the overall TSA accuracy index, the best fit line was: Losses = TSA level 1 accuracy index*-1.192 + 4.706.

4.6. *Low cognitive demand – kills*

There was a significant curvilinear relationship between kills and the TSA accuracy index in the low cognitive demand mission. However, when this relationship was decomposed further to look at the contributions of TSA level 1, 2 and 3 accuracy indices, there was no significant multiple regression solution.

Table 6
Means (M) and standard deviations (SD) of friendly losses and TSA level 1, 2 and 3 accuracy indices at medium cognitive demand, N = 15.

	M	SD
Losses	0.70	0.84
TSA level 1	3.08	0.48
TSA level 2	2.97	0.55
TSA level 3	1.70	0.46

Table 7
Correlations between friendly losses and TSA level 1, 2 and 3 accuracy indices (predictors) and inter-correlations between predictor variables.

	Friendly losses (medium cognitive demand)	TSA level 1	TSA level 2
TSA level 1	−0.81***		
TSA level 2	−0.80***	0.98***	
TSA level 3	−0.48*	0.63**	0.50*

N = 15; *p < 0.05; **p < 0.01; ***p < 0.001.

Table 8
Means (M) and standard deviations (SD) of friendly losses and TSA level 1, 2 and 3 accuracy indices at high cognitive demand, N = 15.

	M	SD
Losses	0.97	1.16
TSA level 1	2.71	0.58
TSA level 2	2.42	0.78
TSA level 3	0.00	0.00

Table 9
Correlations between friendly losses and TSA levels 1, 2 and 3 accuracy indices (predictors) and inter-correlations between predictor variables.

	Losses (high cognitive demand)	TSA level 1	TSA level 2
TSA level 1	−0.72**		
TSA level 2	−0.68*	0.91**	
TSA level 3	0.00	0.00	0.00

N = 15; *p < 0.05; **p < 0.01.

5. Discussion

Fifteen flights each of four F/A-18 fighter aircraft undertook simulated BVR air combat missions which imposed three different levels of cognitive demand required to develop TSA. The cognitive demand was varied by manipulating the datalink-based exchange of information between the aircraft in the flight and the fighter controller (q.v. the approach used by Gorman et al., 2005; Gorman et al., 2006). Inhibiting the free flow of tactical information increased the cognitive demand. Overall, it was observed that as the cognitive demand required to develop and maintain TSA increased, TSA accuracy decreased, the number of friendly losses increased and the number of kills also tended to decrease (see Table 3).

Studies have previously suggested that the relationship between SA and performance can be weak or unclear (e.g., Durso et al., 1998: Endsley et al., 2000; Endsley, 2019; Mansikka et al., 2019a) particularly in a highly dynamic, uncertain environment where success or failure can be a product of chance. Such studies examining the relationship between SA and performance have usually been predicated upon an implicit linear relationship (usually in the form of correlation/regression analyses – see, e.g., Endsley, 2019). R values reported were usually in the region of 0.2–0.5 (e.g., Fracker, 1991; Durso et al., 1998; Strybel et al., 2008; Sulistyawati et al., 2009; Endsley, 2019) but may drop to as low as 0.07 (Strybel et al., 2008). However, in the present study the results showed a strong curvilinear relationship between TSA accuracy and performance (based upon a quadratic regression solution) particularly when performance was measured by avoiding friendly losses. The relationship with performance dependent upon TSA accuracy in some cases accounted for over 85% of the variance. However, the gains in performance decreased disproportionately with increases in TSA accuracy, with the greatest gains in performance observed at lower TSA accuracy. The relationship between TSA accuracy and kills was much weaker and only reached significance in the low cognitive demand condition.

An examination of the strategies employed by the members of the flight provides an explanation for this. At the beginning of each mission, the separation between the flight and the red aircraft provided enough time for the flight to find, fix and track the enemy aircraft with their onboard sensors. Once the enemy aircraft were declared as hostile, the flight attempted to intercept as many red aircraft as possible – often leading to a situation where a single red aircraft was targeted and engaged by more than one flight member. Due to the launch ranges and the speed profiles of the missiles used in the simulation, the flight typically had to defend against the red aircraft's missiles before its own missiles had reached their targets. A flight's most basic defensive manoeuvre was to turn away from the threat and to defeat the incoming missiles kinematically. When the flight re-engaged the remaining red aircraft after such a defensive manoeuvre, the range between the flight and the enemy aircraft was typically much less than it was at the beginning of the mission. Consequently, the flight did not have sufficient time to search again for the red aircraft with their onboard sensors. In low cognitive demand missions, the flight could use the datalink to maintain TSA about the red aircraft even while flying away from them, whereas in high cognitive demand missions the pilots had the start building their radar picture from scratch as they turned towards the red aircraft. The red (simulated enemy) aircraft, however, were programmed with 'perfect SA' and were able to engage the flight as soon as their missile launch ranges permitted. As a result, the cognitive demand manipulation to build TSA had a greater impact on the flight's survivability after their initial defensive manoeuvre – whereas it had little or no effect on the number of red aircraft being intercepted before the first defensive manoeuvre.

These results about the relationship between TSA accuracy and friendly losses complemented those of Sulistyawati et al. (2009) who also found losses to be negatively correlated with TSA when engaging enemy aircraft as a pair (lead and wingman). Furthermore, Fracker (1991) also observed only a weak correlation of SA with kills. Together, these results serve as a reminder that an appreciation of the nature of the task under consideration (in this case an air combat task) is vital when selecting the most appropriate measures of performance. Coupled with the observation that the relationship between (T)SA and performance may best be described as curvilinear rather than linear, this also provides an explanation for the frequently observed dissociation between SA and performance (Durso et al., 1998; Sulistyawati et al., 2009; Mansikka et al., 2019a).

When further decomposing the TSA accuracy indices into TSA level 1, 2 and 3 accuracy indices and examining their relationships with performance, it was observed that in all cases where friendly losses was the metric, all levels of TSA correlated highly with performance in all three TSA demand conditions (low, medium and high). The strongest correlations with performance were at the TSA level 1 (perception) and the weakest at the TSA level 3 (projection) – see Tables 5, 7 and 9. When performance was subject to prediction using multiple regression with the specific TSA level accuracies as predictors, as a result of the high inter-correlations between predictor variables, i.e., TSA levels 1, 2 and 3 accuracy indices, only the TSA 1 level accuracy index entered into the regression equation in all cases. However, this is consistent with Endsley's underpinning theory (Endsley, 1995) as SA level 3 cannot be achieved unless the pilot has SA level 2, and similarly SA level 2 can only be attained if the pilot is already in possession of SA level 1.

The technique for the assessment of TSA developed in the present study had the advantage of not interrupting the progression of the simulated sortie, thus it actually helped to maintain participants' SA as the air combat picture developed and changed. Hence, in addition to virtual air combat simulations, it can also be used in live exercises. (T)SA is dynamic in nature, not static, and is built over a period of time. TSA was also assessed directly from participants and neither inferred from performance, which can be problematic as there is often a dissociation between (T)SA and performance, nor assessed from the observations of subject matter experts. The technique used in this paper acknowledged Endsley's model of SA (Endsley, 1995) and was built upon a well-established knowledge elicitation method, i.e., CDM (see, e.g., Crandall et al., 2006; Plant and Stanton, 2015; Klein and Armstrong, 2004). Even if the assessment of TSA is not the main concern in the debrief of an air combat mission, the proposed technique should serve as a good practice for any debrief by focusing the team's attention to events and activities most relevant to TSA, and it is likely to add value to training simply by promoting reflection and constructive debate during the debrief.

While this study focused in assessing TSA, it also has the potential to evaluate the pilots' mental workload concurrently with the assessment of their SA and the flight's TSA. This type of research is still required to draw a more holistic picture of SA and TSA, and how they are associated with, e.g., performance and mental workload. In addition, the proposed TSA measurement technique requires a skilled interviewer, and the pilots must be motivated to be interviewed. Fortunately, in an air combat training environment, both requirements are typically met.

It was observed that pilots had difficulties in attaining SA level 3 in the high cognitive demand combat mission, see Table 8. There were two primary reasons why it was easy for the flight members to build and maintain level 3 SA in the low cognitive demand condition. First, the flight members were able to almost constantly monitor the red aircraft manoeuvres from their cockpit displays. Second, as the red aircraft followed similar tactics to that the threat force typically uses in any western air combat exercise, it was a simple task for the flight members to anticipate the red aircraft's manoeuvres based on their monitored track history. In the medium cognitive demand condition, building and maintaining SA level 3 was more challenging as only the radio calls and the flight members' onboard sensor tracks contributed to the flight's common tactical picture. Finally, when there was a high cognitive demand required to develop TSA, each flight member had to build and maintain a three-dimensional tactical picture solely from the radio calls and their own onboard sensors. With only two radio channels available for simultaneous receipt and transmission, the frequencies soon became saturated. With the high cognitive demand condition being essentially data limited, the pilots had to conduct highly demanding mental simulation to build and maintain a spatial model of the constantly changing tactical situation. As the level 3 TSA accuracy indices implied, this was a difficult or impossible task even for the qualified fighter pilots.

6. Conclusions

The results of this study indicated that there is a highly significant negative curvilinear, rather than linear, relationship between the TSA accuracy index and friendly losses at all levels of cognitive demand: the gains in performance decreased disproportionately with increases in TSA, with the greatest gains in performance observed with lower TSA accuracy. This relationship was revealed using a new TSA accuracy measurement technique designed for use in simulated air combat introduced and demonstrated in this paper. The technique views TSA as collective knowledge and based TSA accuracy measurement on a recognised knowledge elicitation method, i.e., CDM. By doing so, the technique avoided interrupting the task execution and enabled TSA to be assessed directly from the pilots attending the debriefs.

In conclusion, this paper makes several contributions to ergonomics. First, it identifies attributes which are useful for other TSA studies in the air combat domain. In addition, the way the attributes were developed, will also be helpful when similar attributes are determined in other application domains. Second, it provides an explanation for the frequently observed dissociation between SA and performance in military air missions; the relationship between SA/TSA and performance is likely to be curvilinear, not linear. The greatest performance benefits accrue with the initial gains in TSA. This can also be seen when the levels of TSA are used as predictors of combat performance: the strongest relationship with performance is at TSA level 1. At TSA levels 2 and 3, TSA does not contribute significantly to predicting performance. Third, the introduced TSA accuracy measurement technique will be applicable for TSA accuracy assessment in any domain where the task cannot be interrupted for data collection. In such domains, the new technique could promote and aid TSA assessments carried out in the evaluation of training interventions and the utility of operating procedures, as well as in establishing the competence of teams and assessing the applicability of systems.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Atkinson, R., Shiffrin, R., 1968. Human memory: a proposed system and its control processes. In: Spence, K. (Ed.), The Psychology of Learning and Motivation: Advances in Research and Theory. Academic Press, New York, NY, pp. 89–195.

Bolstad, C., Endsley, M., 2003. Measuring shared and team situation awareness in the army's future objective force. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 47 (3), 369–373. https://doi.org/10.1177/154193120304700325.

Converse, S., Cannon-Bowers, J., Salas, E., 1991. Team member shared mental models: a theory and some methodological issues. Proceedings of the Human Factors Society Annual Meeting 35 (19), 1417–1421. https://doi.org/10.1177/154193129103501917.

Cooke, N., 1994. Varieties of knowledge elicitation techniques. Int. J. Hum. Comput. Stud. 41 (6), 801–849. https://doi.org/10.1006/ijhc.1994.1083.

Cooke, N., Salas, E., Cannon-Bowers, J., Stout, R., 2000. Measuring team knowledge. Hum. Factors 42 (1), 151–173. https://doi.org/10.1518/001872000779656561.

Cooke, N., Stout, R., Salas, E., 1997. Broadening the measurement of situation awareness through cognitive engineering methods. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 41 (1), 215–219. https://doi.org/10.1177/107118139704100149.

Cooke, N., Stout, R., Salas, E., 2001. A knowledge elicitation approach to the measurement of team situation awareness. In: McNeese, M., Endsley, M., Salas, E. (Eds.), New Trends in Cooperative Activities: System Dynamics in Complex Settings. Human Factors, Santa Monica, CA, pp. 114–139.

Cooke, N., Stout, R., Salas, E., 2017. A knowledge elicitation approach to the measurement of team situation awareness. In: Cooke, N., Stout, R., Salas, E. (Eds.), Situation Awareness. Routledge, London, pp. 157–182. https://doi.org/10.4324/9781315087924.

Cooper, G., Harper, R., 1969. The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities, *Report No: NASA TN D-5153*. Moffett Field. National Aeronautics and Space Administration, Ames Research Center, CA.

Crandall, B., Klein, G., Klein, G., Hoffman, R., 2006. Working Minds: A Practitioner's Guide to Cognitive Task Analysis. MIT Press, Cambridge, MA.

Dekker, S., Hollnagel, E., 2004. Human factors and folk models. Cognit. Technol. Work 6 (2), 79–86. https://doi.org/10.1007/s10111-003-0136-9.

Durso, F., Hackworth, C., Truitt, T., Crutchfield, J., Nikolic, D., Manning, C., 1998. Situation awareness as predictor of performance for en-route air traffic controllers. Air Traffic Contr. Q 6 (1), 1–20. https://doi.org/10.2514/ATCQ.6.1.1.

Endsley, M., 1993. A survey of situation awareness requirements in air-to-air combat fighters. Int. J. Aviat. Psychol. 3 (2), 157–168. https://doi.org/10.1207/s15327108ijap0302_5.

Endsley, M., 1995. Toward a theory of situation awareness in dynamic systems. Hum. Factors 37 (1), 32–64. https://doi.org/10.1518/001872095779049543.

Endsley, M., 1998. Situation awareness global assessment technique (SAGAT). Proceedings of the IEEE 1988 National Aerospace and Electronics Conference 3, 789–795. https://doi.org/10.1109/NAECON.1988.195097.

Endsley, M., 2000. Situation models: an avenue to the modeling of mental models. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 44 (1), 61–64. https://doi.org/10.1177/154193120004400117.

Endsley, M., 2015. Situation awareness: misconceptions and misunderstandings. Journal of Cognitive Engineering and Decision Making 9 (3), 4–32. https://doi.org/10.1177/1555343415572631.

Endsley, M., 2019. A systematic review and meta-analysis of direct objective measures of situation awareness: a comparison of SAGAT and SPAM. Hum. Factors 63 (1), 124–150. https://doi.org/10.1177/0018720819875376.

Endsley, M., Garland, D. (Eds.), 1995. Experimental Analysis and Measurement of Situation Awareness. Embry-Riddle Aeronautical University Press, Daytona Beach, FL. Retrieved from. https://apps.dtic.mil/dtic/tr/fulltext/u2/a522540.pdf.

Endsley, M., Jones, W., 2001. A model of inter and intra-team situation awareness: implications for design, training and measurement. In: McNeese, M., Salas, E., Endsley, M. (Eds.), New Trends in Cooperative Activities: Understanding System Dynamics in Complex Environments. Human Factors and Ergonomics Society, Santa Monica, CA, pp. 46–67.

Endsley, M., Sollenberger, R., Stein, E., 2000. Situation awareness: a comparison of measures. In: Endsley, M., Kaber, D. (Eds.), *Proceedings Of the Human Performance, Situation Awareness And Automation: User Centered Design For the New Millennium Conference*, 15–19. SA Technologies Inc, Savannah, GA.

Falkland, E., Wiggins, M., 2019. Cross-task cue utilisation and situational awareness in simulated air traffic control. Appl. Ergon. 74, 24–30. https://doi.org/10.1016/j.apergo.2018.07.015.

Flach, J.M., 1995. Situation awareness: proceed with caution. Hum. Factors 37 (1), 149–157. https://doi.org/10.1518/001872095779049480.

Fiore, S., Ross, K., Jentsch, F., 2012. A Team cognitive readiness framework for small-unit training. Journal of Cognitive Engineering and Decision Making 6 (3), 325–349. https://doi.org/10.1177/1555343412449626.

Fowlkes, J., Lane, N., Salas, E., Franz, T., Oser, R., 1994. Improving the measurement of team performance: the TARGETs methodology. Mil. Psychol. 6 (1), 47–61. https://doi.org/10.1207/s15327876mp0601_3.

Fracker, M., 1991. Measures of Situation Awareness: an Experimental Evaluation. *Report No. AL-TR-1991-0127*. Wright-Patterson Air Force Base. Human Engineering Division, Armstrong Laboratory, OH.

Gilbert, S., 2011. Models-based Science Teaching: Understanding and Using Mental Models. NSTA press, Arlington, VA.

Gilson, R., Garland, D., Koonce, J. (Eds.), 1994. Situation Awareness in Complex Systems. Embry-Riddle Aeronautical University Press, Daytona Beach, FL. Retrieved from. https://apps.dtic.mil/dtic/tr/fulltext/u2/a281448.pdf.

Gorman, J., Cooke, N., Pederson, H., DeJoode, J., 2005. Coordinated awareness of situation by teams (CAST): measuring team situation awareness of a communication glitch. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 49 (3), 274–277. https://doi.org/10.1177/154193120504900313.

Gorman, J., Cooke, N., Winner, J., 2006. Measuring team situation awareness in decentralized command and control environments. Ergonomics 49 (12–13), 1312–1325. https://doi.org/10.1080/00140130600612788.

Hecker, A., 2012. Knowledge beyond the individual? Making sense of a notion of collective knowledge in organization theory. Organ. Stud. 33 (3), 423–445. https://doi.org/10.1177/0170840611433995.

Harris, D., 2011. Human Performance on the Flight Deck. Ashgate Publishing, Ltd, Farnham, UK.

Hoffman, R., Crandall, B., Shadbolt, N., 1998. Use of the critical decision method to elicit expert knowledge: a case study in the methodology of cognitive task analysis. Hum. Factors 40 (2), 254–276. https://doi.org/10.1518/001872098779480442.

Hoffman, R., Shadbolt, N., Burton, A., Klein, G., 1995. Eliciting knowledge from experts: a methodological analysis. Organ. Behav. Hum. Decis. Process. 62 (2), 129–158. https://doi.org/10.1006/obhd.1995.1039.

Joffe, A., Wiggins, M. Cross-task cue utilisation and situational awareness in learning to manage a simulated rail control task. Appl. Ergon., 89, https://doi.org/10.1016/j.apergo.2020.103216.

Johnson, T., O'Connor, D., 2008. Measuring team shared understanding using the analysis-constructed shared mental model methodology. Perform. Improv. Q. 21 (3), 113–134. https://doi.org/10.1002/piq.20034.

Johnson-Laird, P., 1983. Mental Models: towards a Cognitive Science of Language, Inference, and Consciousness. Harvard University Press, Cambridge, MA.

Johnson, T., Lee, Y., Lee, M., O'Connor, D., Khalil, M., Huang, X., 2007. Measuring sharedness of team-related knowledge: design and validation of a shared mental model instrument. Hum. Resour. Dev. Int. 10 (4), 437–454. https://doi.org/10.1080/13678860701723802.

Joint Chief of Staff, 2013. Joint Targeting. Joint Publication, pp. 3–60.

Klein, G., Calderwood, R., Macgregor, D., 1989. Critical decision method for eliciting knowledge. IEEE Transactions on Systems, Man, and Cybernetics 19 (3), 462–472. https://doi.org/10.1109/21.31053.

Klein, G., Armstrong, A., 2004. Critical decision method. In: Stanton, N., Brookhuis, K., Salas, E., Hendrik, H.W. (Eds.), Handbook of Human Factors and Ergonomics Methods. CRC Press, London, UK, pp. 347–356.

Klimoski, R., Mohammed, S., 1994. Team mental model: construct or metaphor? J. Manag. 20, 403–437.

Krampell, M., Solís-Marcos, I., Hjalmdahl, M., 2020. Driving automation state-of-mind: using training to instigate rapid mental model development. Appl. Ergon. 83 https://doi.org/10.1016/j.apergo.2019.102986.

Korean Air Force, 2005. Korean Air Force Tactics, Techniques and Procedures 3-3. Basic Employment Manual F16C. Retrieved from. http://falcon.blu3wolf.com/Docs/Basic-Employment-Manual-F-16C-RoKAF.pdf.

Langan-Fox, J., Code, S., Langfield-Smith, K., 2000. Team mental models: techniques, methods, and analytic approaches. Hum. Factors 42 (2), 242–271. https://doi.org/10.1518/001872000779656534.

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.

Lichacz, F., 2006. An examination of situation awareness and confidence within a distributed information sharing environment. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 50 (3), 344–348. https://doi.org/10.1177/154193120605000328.

Mansikka, H., Virtanen, K., Harris, D., 2019a. Dissociation between mental workload, performance, and task awareness in pilots of high performance aircraft. IEEE Transactions on Human-Machine Systems 49 (1), 1–9. https://doi:10.1109/THMS.2018.2874186.

Mansikka, H., Virtanen, K., Harris, D., Salomäki, J., 2019b. Live–virtual–constructive simulation for testing and evaluation of air combat tactics, techniques, and procedures, Part 1: assessment framework. The Journal of Defense Modeling and Simulation. https://doi.org/10.1177/1548512919886375 ahead of print.

Mansikka, H., Virtanen, K., Harris, D., Salomäki, J., 2019c. Live–virtual–constructive simulation for testing and evaluation of air combat tactics, techniques, and procedures, Part 2: demonstration of the framework. The Journal of Defense

Modeling and Simulation. https://doi.org/10.1177/1548512919886378 ahead of print.

Mansikka, H., Virtanen, K., Harris, D., Jalava, M., 2020. Measurement of team performance in air combat–have we been underperforming? Theor. Issues Ergon. Sci. 1–22. https://doi.org/10.1080/1463922X.2020.1779382.

NATO, 1995. Situation Awareness: Limitations and Enhancement in the Aviation Environment. Canada Communication Group, Quebeck, Canada. Retrieved from. https://apps.dtic.mil/dtic/tr/fulltext/u2/a305000.pdf.

Nguyen, T., Lim, C., Nguyen, N., Gordon-Brown, L., Nahavandi, S., 2019. A review of situation awareness assessment approaches in aviation environments. IEEE Systems Journal 13 (3), 3590–3603. https://doi.org/10.1109/JSYST.2019.2918283.

Plant, K., Stanton, N., 2015. The process of processing: exploring the validity of Neisser's perceptual cycle model with accounts from critical decision-making in the cockpit. Ergonomics 58 (6), 909–923. https://doi.org/10.1080/00140139.2014.991765.

Prince, C., Ellis, E., Brannick, M., Salas, E., 2007. Measurement of team situation awareness in low experience level aviators. Int. J. Aviat. Psychol. 17 (1), 41–57. https://doi.org/10.1080/10508410709336936.

Rajabally, E., Valiusaityte, I., Kalawsky, R., 2009. Aircrew performance measurement during simulated military aircrew training: a review. Proceedings of the AIAA Modeling and Simulation Technologies Conference 5829–5838. https://doi.org/10.2514/6.2009-5829.

Rogers, Y., 1997. A Brief Introduction to Distributed Cognition. Retrieved from. http://yvonnerogers.com/wp-content/uploads/2014/07/dcog-brief-intro.pdf.

Rosenman, E., Dixon, A., Webb, J., Brolliar, S., Golden, S., Jones, K., Sachita, S., Grand, J., Kozlowski, S., Chao, G., Fernandez, R., 2018. A simulation-based approach to measuring team situational awareness in emergency medicine: a multicenter, observational study. Acad. Emerg. Med. 25 (2), 196–204. https://doi.org/10.1111/acem.13257.

Rouse, W., Morris, N., 1986. On looking into the black box: prospects and limits in the search for mental models. Psychol. Bull. 100 (3), 349–363. https://doi.org/10.1037/0033-2909.100.3.349.

Royal Norwegian Air Force, 2001. Royal Norwegian Air Force Tactics, Techniques and Procedures 3-3. *Basic Employment Manual F16C*. Retrieved from: https://www.87th.org/sites/default/files/Downloads/Training/extratraining/Basic%20Employment%20Manual%20AFTTP%203-3%20Vol%205%20.pdf.

Salas, E., Prince, C., Baker, D., Shrestha, L., 1995. Situation awareness in team performance: implications for measurement and training. Hum. Factors 37 (1), 123–136. https://doi.org/10.1518/001872095779049525.

Salas, E., Sims, D., Burke, C., 2005. Is there a big five in teamwork? *Small Group* Res. 36 (5), 555–599. https://doi.org/10.1177/1046496405277134.

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. Theor. Issues Ergon. Sci. 9 (4), 297–323. https://doi.org/10.1080/14639220701561775.

Salmon, P., Stanton, N., Walker, G., Jenkins, D., Ladva, D., Rafferty, L., Young, M., 2009. Measuring Situation Awareness in complex systems: comparison of measures study. Int. J. Ind. Ergon. 39 (3), 490–500. https://doi.org/10.1016/j.ergon.2008.10.010.

Stanton, N., Salmon, P., Walker, G., Salas, E., Hancock, P., 2017. State-of-science: situation awareness in individuals, teams and systems. Ergonomics 60 (4), 449–466. https://doi.org/10.1080/00140139.2017.1278796.

Stanton, N., Stewart, R., Harris, D., Houghton, R., Baber, C., McMaster, R., Salmon, P., Hoyle, G., Walker, G., Young, M., Linsell, M., Dymott, R., Green, D., 2006. Distributed situation awareness in dynamic systems: theoretical development and application of an ergonomics methodology. Ergonomics (49), 1288–1311. https://doi.org/10.1080/00140130600612762, 12-13.

Strybel, T., Vu, K., Kraft, J., Minakata, K., 2008. Assessing the situation awareness of pilots engaged in self spacing. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 52 (1), 11–15. https://doi.org/10.1177/154193120805200104.

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.

Svenmarckt, P., Dekker, S., 2003. Decision support in fighter aircraft: from expert systems to cognitive modelling. Behav. Inf. Technol. 22 (3), 175–184. https://doi.org/10.1080/0144929031000109755.

van der Haar, S., Segers, M., Jehn, K., Van den Bossche, P., 2015. Investigating the relation between team learning and the team situation model. *Small Group Res.* 46 (1), 50–82. https://doi.org/10.1177/1046496414558840.

Vidulich, M., Dominguez, C., Vogel, E., McMillan, G. (Eds.), 1994. Situation Awareness: Papers and Annotated Bibliography. Wright-Patterson Air Force Base. Armstrong Laboratory, Human Engineering Division, OH. Report No. AL/CF-TR-1994-0085.

Waag, W., Houck, M., 1994. Tools for assessing situational awareness in an operational fighter environment. Aviat Space Environ. Med. 65 (5, Sect 2, Suppl. l), A13–A19.

Wegner, D., 1985. A computer network model of human transactive memory. Soc. Cognit. 13 (3), 319–339. https://doi.org/10.1521/soco.1995.13.3.319.

Weigl, M., Catchpole, K., Wehler, M., Schneider, A., 2020. Workflow disruptions and provider situation awareness in acute care: an observational study with emergency department physicians and nurses. Appl. Ergon. 88 https://doi.org/10.1016/j.apergo.2020.103155.

Wellens, A., 1993. Group situation awareness and distributed decision making: from military to civilian applications. In: Castellan Jr., N.J. (Ed.), Individual and Group Decision Making: Current Issues. Lawrence Erlbaum Associates Inc, Hillsdale, NJ, pp. 267–291. https://doi.org/10.4324/9780203772744.

Wickens, C., 1991. Processing resources and attention. In: Damos, D. (Ed.), Multiple-task Performance. Taylor and Francis, London, UK, pp. 3–34.

Wickens, C., Lee, J., Liu, Y., Becker, G., 2004. An Introduction to Human Factors Engineering. Pearson Education, Upper Saddle River, NJ.

Wickens, C., 2008. Situation awareness: review of Mica Endsley's 1995 articles on situation awareness theory and measurement. Hum. Factors 50 (3), 397–403. https://doi.org/10.1177/1934578X1801300212.

Wilson, J., Rutherford, A., 1989. Mental models: theory and application in human factors. Hum. Factors 31 (6), 617–634. https://doi.org/10.1177/001872088903100601.