International Journal of Industrial Ergonomics 86 (2021) 103233

Contents lists available at ScienceDirect

International Journal of Industrial Ergonomics

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

Investigating the validity of subjective workload rating (NASA TLX) and subjective situation awareness rating (SART) for cognitively complex human–machine work

Per Øivind Braarud

Institute for Energy Technology/OECD Halden Reactor Project, PB 173, NO-1751, Halden, Norway

ARTICLE INFO

Keywords:
Mental Workload
Situation Awareness
Subjective Measure
NASA TLX
SART
Control Room
Human-System Evaluation

ABSTRACT

Subjective workload and situation awareness measures, such as the NASA task load index (TLX) and the situational awareness rating technique (SART), are frequently used in human–system evaluation. However, the interpretation of these ratings is debated. In this study, empirical evidence for the measures' theoretical assumptions was investigated by comparing operators' ratings collected immediately after performing a scenario and ratings collected after operators' acquisition through a video review of the scenario, knowledge of actual system states. Eighteen licensed control room operators participated in the simulator study, running 12 relatively challenging scenarios. It was found that the interpretation of TLX items involving introspection remained stable after operators acquired factual scenario knowledge, while the interpretation of items involving the perception of external events, such as situation awareness and performance, depended on the operators' scenario knowledge. The result shows that operators' ratings could discriminate between mental effort, performance, frustration, and situation awareness. No clear evidence for the SART index as a measure of situation awareness was found. Instead, a subjective situation awareness measure developed for this study was distinct from workload and related to operator performance, showing that this type of measure warrants future investigation of its validity. The study findings help in developing measurement procedures and interpreting subjective measures. Finally, the study reveals that informing operators about the scenario can provide useful subjective ratings of situation awareness and performance. Future research should include procedures for how to inform participants adequately and efficiently in subjective assessments.

1. Introduction

Mental workload and situation awareness are important criteria for designing and assessing human–machine systems (Endsley et al., 1998; Salmon et al., 2009; O'Hara et al., 2012). Workload and situation awareness measures are applied to obtain sensitive evaluations and gain an in-depth understanding of complex task performance (O'Donnel and Eggemeier, 1986; Endsley et al., 1998). Important questions include how human–system design and organisation of work influence workload and situation awareness, and to what extent a given human–machine configuration represents optimal or acceptable levels (Reid and Colle, 1988; Endsley, 2000a; Young et al., 2015). To adequately address these questions, measurement credibility and utility are important considerations (Muckler and Seven, 1992; Annett, 2002).

Mental workload and situation awareness are frequently viewed as separate but interrelated constructs (Endsley, 2000a; Vidulich and Tsang, 2015). For example, when the attentional resources involved in performance compete with the resources needed for monitoring and comprehension, situation awareness may decrease (Vidulcih and Tsang, 2015). However, the operator's increased effort may be related to increased situation awareness. By improving the human-system interface or increasing the level of expertise, one could reduce the workload and increase situation awareness—a frequent goal of system design efforts (Vidulich, 2000; Endsley, 2000a; Vidulich and Tsang, 2015). It follows that how workload and situation awareness relate to each other and under what conditions may provide important insights regarding human–system safety and efficiency (Vidulich and Tsang, 2015). Consequently, separate and valid measures of each construct are warranted (Endsley, 2000a; Parasuraman et al., 2008).

Subjective assessment techniques are frequently used due to their ease of use, low cost of application, and sensitivity to varying conditions (Reid and Nygren, 1988). The NASA task load index (TLX) (Hart and

E-mail address: per.oivind.braarud@ife.no.

https://doi.org/10.1016/j.ergon.2021.103233
Received 27 May 2021; Received in revised form 28 September 2021; Accepted 26 October 2021
Available online 2 November 2021

Staveland, 1988) is the most popular subjective workload measure (De Winter, 2014; Grier, 2015), and the situational awareness rating technique (SART) (Taylor, 1990) is the most widely used subjective situation awareness measure (Endsley et al., 1998; Salmon et al., 2009). However, subjective workload techniques have been questioned due to a lack of correspondence with performance-based and physiological workload measures (Yeh and Wickens, 1988; Matthews et al., 2020), and subjective situation awareness has been found to dissociate with objective measures (Endsley, 2020). Therefore, it is important to improve our knowledge of what subjective measures can inform us (Messick, 1990, 1995). Subjective ratings concern the assessment of events in the external world and internal sensations and feelings (Annett, 2002). The latter is obtained through introspection, while events in the external world are objects of direct verification or consensus view. Some external events are directly verifiable (Muckler and Seven, 1992). For example, if a system had stopped, the time available for the task was 5 min. Some events are verifiable by consensus; for example, the event created many alarms in need of attention. The situation implied that the system needed to be shut down. Internal sensations are hard to verify, but mental workload, as an example, can considerably be inferred from external conditions and performance (Colle and Reid, 2005), psychophysiological response (Charles &Nixon, 2019), and overt behaviour (Braarud et al., 2020; Gan et al., 2020).

Mental workload likely involves both private sensations regarding effort and verifiable elements, such as task assessment (Annett, 2002). Situation awareness involves the perception of events in the external world and is probably the object of direct verification and consensus (Parasuraman et al., 2008). Therefore, it is interesting to note that subjective ratings are frequently collected immediately after completing a scenario (post-session) without the participant having explicit information about the scenario's operational challenges, actual system states, or the consequences of one's actions. Ratings are based on the operator's memory recall of their observations during the performance. This is contrary to expert observers who are usually fully informed about these issues (Endsley, 2020). Thus, the operator may lack an adequate point of reference for the assessment. Could we better understand the meaning of subjective ratings by providing the operator with an external reference to what actually happened during the scenario?

Validity and reliability are critical considerations when selecting human factors measures and interpreting their results (Annett, 2002; Salmon et al., 2009). It is important that the measures represent the phenomena one attempts to investigate and that the measures collected can be interpreted according to the purpose of the evaluation. From a psychometric perspective (Murphy and Davidshofer, 1994; Messick, 1995; Annett, 2002), as with many human factors phenomena, subjective mental workload and subjective situation awareness are constructs. Construct validity is investigated through a process of collecting evidence for or against the accuracy of interpretations and actions taken based on the measurement (Messick, 1990, 1995). Validity evidence includes a review of the measurement content, the internal structure of its components, and relationships with phenomena external to the measure (Messick, 1990; Murphy and Davidshofer, 1994). Experts can assess the extent to which the measurement items represent the concept (Fracker, 1991; Salmon et al., 2009). One can compare the structure identified by factor analysis to the theoretically proposed structure (Annett, 2002). Also, sensitivity to varying loads and alternative human–machine configurations are important applied measurement criteria (O'Donnel and Eggemeier, 1986; Endsley, 2000b; Salmon et al., 2009).

Hart and Staveland (1988) developed the NASA TLX. The measure was developed based on substantial theory and considerable empirical testing (Hart, 2006; De Winter, 2014). Using the measure has extended to contexts exceeding its empirical developmental basis, for example, air traffic control, process control, healthcare, and military (Hart, 2006; Grier, 2015). Hart and Staveland (1988, p. 144) defined workload as "… the cost incurred by human operators to achieve a specific level of performance." They focused on three aspects, with the following measurement items: a) The external demand imposed by the tasks—three items consider mental, physical, and temporal demands; b) Effort based on the operator perception of task demand, including the self-regulation of effort and understanding of task demand based on perceived performance—the items effort and performance cover these aspects; c) psychological impact of perceived task demand, effort, and performance—captured by the item frustration. Hart and Staveland (1988) also referred to these three aspects as task, behaviour, and subject related, respectively. The intended application, which is frequently used, occurs immediately after completing the task or scenario (Hart and Staveland, 1988). From the development of Hart and Staveland's (1988) measure, one can develop several assumptions about the NASA TLX. (a) The rating of demand, the work loaded on the operator, could be substantially influenced by the operator's perception of its significance and magnitude, including what system behaviour is detected and understood during a scenario. Effort or resources invested, however, can be viewed as representing introspective characteristics. One can hypothesise that being informed about what actually happened in a scenario, e.g., system and component states, operational consequences of one's own and team members' performance, could influence the understanding of demand to a higher extent than this information would influence perception of effort invested. (b) TLX performance should reflect self-regulation and should therefore be related to effort. (c) TLX performance should be related to frustration, and eventually, this relationship should be modified by task demand. According to Hart and Staveland (1988, p. 166), frustration provides "… information about how comfortable operators felt about the effectiveness of their efforts relative to the magnitude of the task demands imposed on them." (d) Adding to frustration relating to performance, frustration represents the psychological impact of perceived task demand and effort, and one can hypothesise that frustration should relate to all TLX items. This type of relationship was found by Hart and Staveland (1988)—that their preliminary scales of stress and frustration were highly correlated with any other subscales.

Tayler's (1990, p. 3–3) working definition of situation awareness when developing SART is that "Situational Awareness is the knowledge, cognition and anticipation of events, factors and variables affecting the safe, expedient and effective conduct of the mission". The research behind the SART development was influenced by the workload paradigm with its aim of optimising operator workload (Taylor, 1990). Knowledge elicitation and structural analysis resulted in three broad dimensions (Taylor, 1990, pp. 3–7): (a) Demands on Attentional Resources (Instability, Complexity, Variability), (b) Supply of Attentional Resources (Arousal, Concentration, Division of Attention, Spare Capacity), and (c) Understanding of the Situation (Information Quantity, Information Quality, Familiarity). Several studies have found that SART is substantially correlated with workload (Hendy, 1995; Selcon et al., 1991; Loft et al., 2015). This is not very surprising given SART's developmental basis, and the development procedure applied by Taylor (1990). The knowledge elicitation technique generated scenarios representing low and high situation awareness. The scenarios correspondingly varied in workload. For example, "Flying in formation in an unfamiliar aircraft working at the limit of your capacity" vs. "Approaching to land in good weather at a familiar airfield, in a familiar aircraft fitted with good displays". Consequently, constructs elicited from subjects tended to describe the task demand, such as attentional demand, familiarity, and complexity—and constructs that one could relate directly to workload, such as spare capacity, workload, and arousal (Taylor, 1990, Table 1). Taylor (1990, p. 3–11) suggested that situation awareness can be enhanced by controlling the demand on attentional resources and improving the supply of attentional resources, for example, by prioritising and cuing tasks or exploiting mental resource modalities. Taylor and Selcon (1994) developed a formula for the SART index as SA = Understanding—(Demand – Supply). As the formula prescribes, an imbalance between demand and supply should increase or reduce SA beyond what is measured by the understanding

Table 1
Post-Session and Post-Video means and 95% confidence intervals of subjective measures for each control room position and total. (RO = Reactor Operator, TO = Turbine Operator, SS = Shift Supervisor, Team Average = Average of RO, TO and SS).

		Post-Session				Post-Video			
		RO	TO	SS	Team Average	RO	TO	SS	Team Average
NASA TLX	Mental	7.04	5.88	6.13	6.35	6.79	5.83	6.04	6.22
		[6.53, 7.56]	[5.50, 6.25]	[5.58, 6.67]	[6.06, 6.63]	[6.29, 7.30]	[5.47, 6.20]	[5.48, 6.60]	[5.94, 6.50]
	Physical	4.78	3.04	2.19	3.34	4.78	3.32	2.18	3.43
		[4.26, 5.29]	[2.62, 3.47]	[1.94, 2.45]	[3.06, 3.61]	[4.28, 5.28]	[2.86, 3.78]	[1.95, 2.41]	[3.15, 3.70]
	Temporal	6.19	4.64	5.78	5.54	6.10	4.81	5.76	5.56
		[5.67, 6.71]	[4.20, 5.08]	[5.25, 6.31]	[5.24, 5.83]	[5.55, 6.64]	[4.36, 5.25]	[5.24, 6.29]	[5.26, 5.85]
	Perfor-mance	7.35	6.74	8.11	7.40	7.11	6.25	8.04	7.13
		[6.90, 7.79]	[6.33, 7.14]	[7.71, 8.52]	[7.15, 7.65]	[6.60, 7.62]	[5.78, 6.72]	[7.63, 8.45]	[6.85, 7.42]
	Effort	6.44	5.31	5.89	5.88	6.24	5.26	5.33	5.61
		[5.92, 6.97]	[4.93, 5.68]	[5.36, 6.42]	[5.60, 6.16]	[5.69, 6.79]	[4.89, 5.64]	[4.80, 5.87]	[5.32, 5.90]
	Frustration	4.68	4.50	2.96	4.05	4.33	4.50	3.07	3.97
		[4.23, 5.13]	[3.98, 5.02]	[2.52, 3.40]	[3.76, 4.33]	[3.88, 4.79]	[4.00, 5.00]	[2.64, 3.50]	[3.69, 4.24]
SART	Demand	6.72	5.74	6.86	6.44	6.39	5.76	6.74	6.30
		[6.29, 7.16]	[5.33, 6.15]	[6.35, 7.37]	[6.17,6.71]	[5.93, 6.85]	[5.40, 6.13]	[6.25, 7.22]	[6.04, 6.55]
	Supply	7.18	5.61	6.63	6.47	6.85	5.51	6.44	6.27
		[6.78, 7.58]	[5.20, 6.02]	[6.14, 7.11]	[6.21, 6.73]	[6.46, 7.23]	[5.13, 5.90]	[5.96, 6.93]	[6.02, 6.52]
	Under-standing	7.46	7.17	7.82	7.48	7.33	6.82	7.83	7.33
		[7.06, 7.86]	[6.75, 7.58]	[7.47, 8.17]	[7.26, 7.71]	[6.91, 7.75]	[6.36, 7.28]	[7.45, 8.21]	[7.08, 7.57]
SA3	Observing	7.42	7.29	8.28	7.66	7.25	6.78	7.64	7.22
		[7.01, 7.83]	[6.94, 7.64]	[7.88, 8.68]	[7.43, 7.89]	[6.78, 7.72]	[6.32, 7.23]	[7.09, 8.18]	[6.94, 7.51]
	Compre-hension	7.78	7.57	8.49	7.94	7.64	7.25	8.31	7.73
		[7.41, 8.15]	[7.16, 7.98]	[8.20, 8.77]	[7.73, 8.16]	[7.21, 8.07]	[6.81, 7.69]	[7.92, 8.69]	[7.49, 7.98]
	Predict	7.85	7.50	7.92	7.75	7.72	7.29	8.15	7.72
		[7.50, 8.19]	[7.15, 7.85]	[7.51, 8.32]	[7.55,7.96]	[7.31, 8.13]	[6.92, 7.66]	[7.77, 8.53]	[7.50, 7.95]
	TLX	5.63	4.77	4.47	4.96	5.52	4.91	4.39	4.94
	Index	[5.25, 6.01]	[4.45, 5.09]	[4.12, 4.82]	[4.75, 5.17]	[5.14, 5.90]	[4.59, 5.23]	[4.06, 4.73]	[4.73, 5.15]
	SART	7.92	7.04	7.58	7.51	7.79	6.57	7.54	7.30
	Index	[7.50, 8.34]	[6.51, 7.58]	[7.12, 8.04]	[7.24, 7.79]	[7.27, 8.31]	[6.07, 7.07]	[7.14, 7.95]	[7.02, 7.58]
	SA3	7.68	7.45	8.23	7.79	7.54	7.11	8.03	7.56
	Index	[7.33, 8.03]	[7.11, 7.79]	[7.92, 8.54]	[7.59, 7.98]	[7.12, 7.95]	[6.71, 7.50]	[7.66, 8.41]	[7.33, 7.79]

element. For example, supply exceeding demand would increase situation awareness beyond an operators' understanding. From the SART basis, one can develop assumptions of a) demand and supply of attention mainly capturing workload; (b) Similar to the NASA TLX dimensions, demand can be seen as operator perception of the external task, while supply of cognitive resources tends toward a subject-oriented element susceptible to introspection. These could be expected to behave similar to NASA TLX workload items depending on the operator being informed about what actually happened in a scenario; c) One can also assume that the element demand–supply of the SART formula, representing a factor influencing situation awareness, should be related to the rating of understanding.

Since it is debateable to what extent the SART measure covers situation awareness, workload, or their relationship, the study found it imperative to consider a supplemental theoretical basis for situation awareness. Endsley's three-level theory (1995a; 1995b) is the most popular and probably the most cited theory of situational awareness (Salmon et al., 2009). Endsley (1995a, p. 36) defined situation awareness as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future''. Situation awareness Level 1 concerns the observation of the status and behaviour of, for example, system parameters, alarms, and functionality. Level 2 concerns the understanding of the observations for managing and controlling the system, while Level 3 concerns the anticipation and prediction of future system development. The theory is the basis for the Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1995b), an objective freeze-probe-based measure of situation awareness. However, to investigate subjective measures, this study developed a simple subjective measure based on Endsley's theory. Similar subjective measures have been developed and applied to aviation and military command and control (McGuinness and Foy, 2000; Matthews and Beal, 2002; McGuinness, 2021). From the theoretical basis and previous studies (Endsley et al., 1998), the assumptions include a) that a subjective measure based on Endsley's theory would be distinct from workload, b) being a pure situation awareness measure, it should be closer related to operator performance than the SART, and c) operator ratings could distinguish the theories' three levels from each other.

This study set out to provide evidence related to the meaning of scores from the most popular post-session subjective measures of workload and situation awareness. Based on the measures' theoretical basis, it was hypothesised that operator assessment of workload and situation awareness items related to an external reference would be influenced by knowledge about the scenarios' intended task demand, actual scenario development, and knowledge of own performance. However, assessment based on introspection was expected to be minimally influenced by this type of knowledge. In addition, due to the SART measure integrating situation awareness and workload, the purpose was to explore a simple subjective measure of situation awareness based on Endsley's (1995a) theory while investigating the demand–supply element of the SART formula. The research questions were studied by collecting nuclear operators' subjective ratings immediately after performing scenarios in a full-scope research simulator (post-session, "non-informed" assessment) and collecting the same subjective ratings after operators were informed about the scenario demands and completing a scenario replay/video analysis of the scenario (post-video, "informed" assessment).

2. Method

2.1. Participants

Eighteen licensed operators from a Nordic nuclear power plant, organised as six three-person teams, participated in the study. Each team comprised a supervisor, a reactor operator, and a turbine operator. Their mean age was 39.2 years (SD = 11.9), ranging from 26 to 62 years, and their mean control room work experience was 10.6 years (SD = 10.8), ranging from 1 to 35 years. Five of the six teams comprised operators working as a team at their home plant, while one team was assembled from two different home plant teams. All members of a given team worked at the same reactor unit and thereby possessed shared competence in technical work, collaboration practices, and communication procedures. Operators maintained their competence by regular simulator-based training at their home plant. In this study, the teams were instructed to use teamwork practices and operating procedures as they ordinarily would in their daily work and in their home plant training. The study was reviewed and approved by the Halden Reactor Project Human Studies Review Committee and was performed according to the Halden Reactor Project's human participant protection procedures.

2.2. Measures

2.2.1. NASA TLX

The study utilised an unweighted version of the NASA TLX (Hart and Staveland, 1988), often referred to as the raw TLX (Byers et al., 1989; Nygren, 1991). The NASA-TLX comprises the following six items: mental demand, physical demand, temporal demand, performance, effort, and frustration. All questions, except the performance question, offered a scale ranging from "very low" to "very high". The rating scale for the performance question ranged from "perfect" to "failure". Each question's scale ranged from 1 to 11.

2.2.2. Situation awareness rating technique (SART)

The study used the 3-item version of the SART (Taylor, 1990). This version is often referred to as 3D SART. The measure comprised the following dimensions: Demand—demands on attentional resources, Supply—supply of attentional resources, and Understanding—understanding of the situation. The rating scale for each question ranged from 1 to 11. The items were worded, and scale endpoints labelled as follows in parentheses; *The situation was* (Very stable, Simple and straight forward, Few variables changing—Unstable, changes suddenly, Many interrelated components, Many variables changing); *Attention, my effort was* (Low alertness, Focused on one aspect, Much spare capacity—High alertness, Concentrating on many aspects, No spare capacity); *My understanding of the situation was* (Fully informed and full understanding, Very familiar situation—Very limited informed/understanding, Very novel situation).

2.2.3. Subjective situation awareness three levels (SA3)

A self-rating measure based on Endsley's theory of situation awareness was developed specifically for this study. The measure was given the preliminary label "SA3". Starting from the definition "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future'' (Endsley, 1995a, p. 36), three items were developed to represent each of the three levels of situation awareness. The items were worded, and scale endpoints labelled as follows in parentheses; *My observation of critical information* (Identified all needed information—Missed important information); *My understanding of what was going on* (Fully understood—Did not make sense to me); *I could look ahead, and foresee what was going to happen* (Very accurately—Could not predict). The rating scale for each question ranged from 1 to 11. The SA3 measure was similar to the subjective measures developed by McGuinness and Foy (2000) and Matthews and Beal (2002). However, beyond Endsley's three-level theory, these two measures included a workload subscale, and items about synthesising situation awareness with one's course of action.

2.2.4. SCORE performance evaluation

The performance assessment used the Supervisory Control and Resilience Evaluation (SCORE) framework (Braarud et al., 2015, 2016). Using the framework, a subject matter expert developed a task-specific assessment sheet for each scenario. The assessment items developed considered the control room team's monitoring, interpretation, strategy, action, verification, teamwork, work process, and goals. During operators' review of their own team performance, utilising the scenario replay tool described below, each operator individually evaluated each item regarding the degree of acceptability on a rating scale. The scale ranged from 1 to 6, where 1–2 defined levels of unacceptable and 5–6 defined levels of acceptable. The middle values of 3 and 4 represented borderline acceptability. To create a scenario index score for each operator, an average was calculated for the task-specific items belonging to each of the control room positions. Consequently, the calculation resulted in a task-specific performance index for each operator for each scenario, ranging between 1 and 6.

2.3. Scenarios

Each team participated in 12 scenarios designed to last for about twenty to thirty min. The scenarios were designed by a subject matter expert with 20 years' experience in scenario design and operator performance assessment at nuclear training and research simulators. The subject matter expert addressed the study purpose by utilising the experiences from numerous human factors experiments previously performed at the research site (Braarud and Kirwan, 2011; Øvre, 2011; Braarud and Svengren, 2020). The resulting scenarios were jointly reviewed by the subject matter expert and a human factors researcher (the author) utilising the full-scope research simulator. The general scenario structure comprises initial work, normal operation, or periodic testing, including minor plant system failures, and aggravating deviations inducing actuation of plant safety functions (reactor protection systems). To make parts of the scenarios complicated and cognitively demanding, malfunctions were simulated for several safety system components. Malfunctions included failed instrumentation, spurious actuated safety functions, and complicated plant system status due to the loss of external and internal power. Consequently, the scenario tasks posed the highest load on the control room team's reactor operator, although the turbine operator was also substantially involved in verifying and controlling plant safety. The supervisor had the overall task of overviewing plant safety, deciding strategy, and supervising the team's work. The scenarios included events commonly found in nuclear power plant safety analysis, such as loss of reactor coolant accidents, loss of all offsite power, loss of turbine condenser, and loss of main feedwater. While individual failures like these are included in the operators' regular home plant training, combining the main event and multiple safety component malfunctions made the specific combinations of malfunctions relatively unfamiliar to the operators. The team applied event-based operating procedures to acquire an overview of the plant's safety status and to develop a basis for selecting a strategy to mitigate the situation and control plant safety. The scenarios were counterbalanced using the Latin-square procedure described in Kirk (1995).

2.4. Simulator and session recording

The study was performed in IFE's Halden Human-Machine Laboratory (IFE, 2021), which is a full-scope research simulator based on an advanced nuclear power plant. The simulator has a fully computerised human–machine interface. Fig. 1 shows the control room layout. The shift supervisor workstation is at the back (closest to the camera), the reactor operator workstation is to the left, and the turbine operator workstation is to the right. The large screen display at the front provides a plant overview.

The simulator sessions were recorded with the laboratory's Video Audio Data Analysing (VAD) tool for use in post-video assessment. Each operator wore a headset with a microphone. The tool provided synchronised play of simulator logs, video, and audio from a scenario completed in the simulator. The recording included the simulated

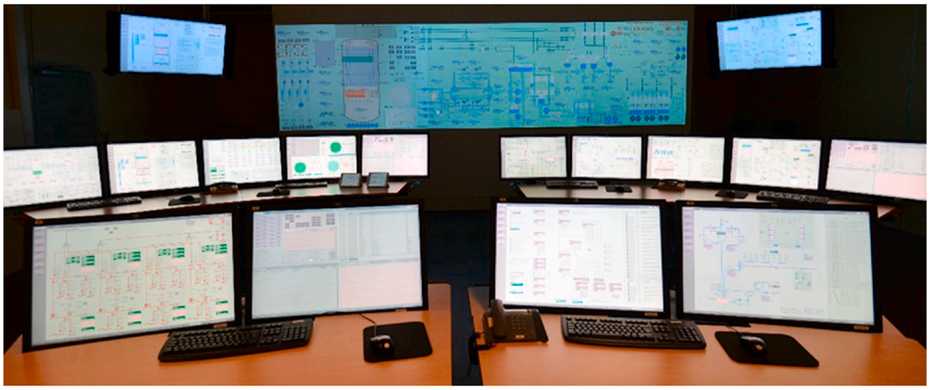

Fig. 1. The Halden Human–Machine Laboratory control room.

plant's process development (alarms, process parameters, and process events), operator process commands, navigation and interfaces accessed by the operator, a video of each operator workstation alongside an overview video of the control room, and separate audio recordings from each of the control room operators. The operator could play, pause, rewind, and forward the scenario during the performance review.

2.5. Study design and study procedure

The NASA TLX, SART, and SA3 rating questionnaires, the performance assessment, and the scenario replay tool were explained to the participating operators before performing the scenarios. The explanation included demonstrating the rating questionnaires, the scenario replay, and the performance assessment tool. Just after completing the scenario, the rating questionnaires were administered. Operators answered the rating questionnaires individually. Thereafter followed a short team briefing of the scenario's task demand, including a brief explanation of the plant failures implemented during the scenario and their consequences to plant operation. No discussions between team members were allowed during this briefing. After the briefing, operators individually, at separate workstations, utilised the scenario replay tool and performed the SCORE performance assessment for the scenario just performed. Operators could, at their own pace, play, pause, rewind, and forward the scenario during the performance assessment. There was no time limit for the assessment. Laboratory staff were, upon request from the operators, available for assistance on the technical aspects of answering the rating questionnaires, the scenario replay tool, or the performance assessment tool. After completing the performance assessment, each operator individually performed a second rating of the subjective measures of NASA TLX, SART, and SA3. The sequence of activities is depicted in Fig. 2, and this sequence was repeated for each of the 12 scenarios.

3. Results

3.1. Subjective measures descriptive data

Table 1 shows the mean and 95% confidence interval for each measurement item for the post-session condition and the post video condition, for each control room position and the tam average. The lower part of the table shows three indexes—the average of the six TLX items, SART calculated according to its formula (SA = U – (D−S)), and the average of the three SA3 items.

Table 1 shows that the post-session and post-video mean of the ratings and the 95% confidence intervals were not very different. Noteworthy differences were the slightly lower team average of NASA TLX performance and effort, and SA3 observing in the post-video condition, differences in team average of 0.26, 0.27 and 0.44, respectively. Table 1 also shows that the operators' ratings of the TLX dimensions varied substantially. Looking at the post-session ratings, the team average ranged from a physical demand of 3.34 to a performance of 7.40. The team average SART understanding was about 1 scale point above demand and supply. The TLX index was substantially lower than the SART and SA3 indexes for all three control room positions and the team average.

3.2. Reliability in terms of Cronbach's alpha

Reliability, in terms of internal consistency, was assessed with Cronbach's alpha. The internal consistencies of the scales were high, and remarkably so for the SA3. The α values are as follows: NASA TLX post-session (6 items; α = 0.86), NASA TLX post-video (6 items; α = 0.83), SART post-session (3 items; α = 0.74), SART post-video (3 items; α = 0.71), SA3 post-session (3 items; α = 0.88), and SA3 post-video (3 items; α = 0.90). To compare the reliability of the three item SART and SA3 scales with the NASA TLX six item scale, an alpha was calculated based on the assumption of adding three more items, assuming the same intercorrelations as the three original items. The resulting α values are as follows: SART post-session (assumed 6 items; α = 0.85), SART post-

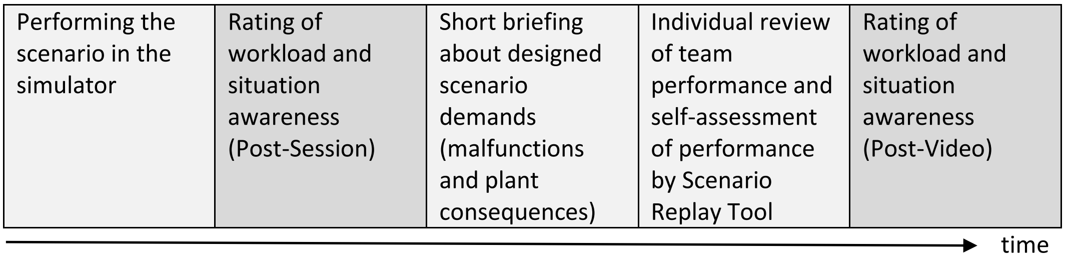
Performing the scenario in the simulator
Rating of workload and situation awareness (Post-Session)
Short briefing about designed scenario demands (malfunctions and plant consequences)
Individual review of team performance and self-assessment of performance by Scenario Replay Tool
Rating of workload and situation awareness (Post-Video)
time

Fig. 2. Sequence of study activities for each scenario.

video (assumed 6 items; α = 0.83), SA3 post-session (assumed 6 items; α = 0.94), and SA3 post-video assumed 6 items; α = 0.95). Assuming six items, the results suggest similar high reliability of the NASA TLX and the SART, while SA3 showed very high reliability.

3.3. *Item correlation and construct structure*

Table 2 shows the bi-variate correlations between the NASA TLX, SART, and SA3 items for the operators' post-session and the post-video rating.

Table 2 shows a similar pattern of bi-variate correlation between the items of the post-session rating and the post-video rating. Some noteworthy observations are high bi-variate correlations between the TLX dimensions of mental demand, temporal demand, and effort. Also, the SART items demand and supply correlated substantially with the above mentioned TLX items. The SA3 items were highly intercorrelated, while only the SART item understanding correlated with the SA3 items. The TLX performance item correlated positively with the SA3 items and the SART understanding item, while performance and frustration were substantially negatively correlated.

Table 3 shows the correlation between post-session and post-video ratings for each item of the NASA TLX, SART, and the SA3 measures. As hypothesised, the correlation for effort and frustration was relatively high, while the correlation for TLX performance, SART understanding, and SA3 items were relatively low. Surprisingly, unlike the assumptions based on the NASA TLX theoretical basis, the correlations for mental demand, physical demand, and temporal demand were relatively high.

A factor analysis of the post-session and post-video ratings together revealed an interesting structure from the operator ratings. The interpretation of the resulting scree plot alongside the criterion of eigenvalues exceeding 1 (Kim and Mueller, 1978) suggested four underlying factors. The principal axis method for factor extraction was applied, and normalised varimax rotation was applied for interpreting the factors. Table 4 shows the resulting factor loadings of the items.

Factor 1 was interpreted as a workload dimension stable across both the post-session and post-video ratings. The TLX items mental demand, temporal demand, and effort alongside the SART items demand and supply, from both post-session rating and post-video rating, loaded on this factor. Interestingly, regarding the SA3 items and SART understanding, the post-session and post-video ratings loaded on different factors. Factor 2 was interpreted as an "informed" situation awareness dimensions, while Factor 4 was interpreted as a "non-informed" situation awareness dimension. NASA TLX performance post-session was loaded mostly on the non-informed SA dimension, and NASA TLX post-video was loaded mostly on the informed SA dimension. These TLX performance loadings supported the interpretation of an informed and a non-informed SA dimension. Factor 3 was defined by the high loadings from the TLX physical demand rating both post-session and post-video. TLX frustration, both post-session and post-video, loaded moderately on this factor. Factor 3 was interpreted as a physical demand factor.

3.4. *The SART Demand–Supply element*

Investigating the bi-variate correlations and factor analysis suggested that the element D–S (Demand–Supply) was not systematically related to the SART item understanding or to the NASA TLX or SA3 items. The correlation between D–S and understanding was 0.10 (n.s.) and 0.13 (n.s.) for the post-session rating and the post-video rating, respectively. The correlation between the D–S element and SA3 ranged from 0.04 to 0.10 and from 0.06 to 0.09 for the post-session rating and the post-video rating, respectively, and the correlation between the D–S and TLX items ranged from 0.08 to 0.31 and from 0.10 to 0.24 for the post-session and the post-video ratings, respectively. Replacing the SART items demand and supply, both post-session and post-video, with the respective element D–S (Demand–Supply) in the factor analysis above resulted in no clear factor loadings. The loadings on the four factors ranged from 0.003 to 0.17.

3.5. *Sensitivity*

Sensitivity of the measures to varying loads for the positions of a control room team was measured by analysis of variance (ANOVA). The scenarios were designed with a predominance of malfunctions and operational challenges for the reactor side of the plant, thereby creating the highest load for the reactor operator. An overall analysis was performed for each index measure, and the effects of team position, and conditions of rating (post-session vs post-video) were investigated by 3X2 ANOVA. The mean ratings and 95% confidence intervals for the two factors are illustrated in Fig. 3.

For the NASA TLX, the effect of operator position was significant, F (2,213) = 11.65, p < .001, while there was no significant difference in rating post-session versus post-video, F(1,213) = 0.12, p = .73. Tukey's

Table 2
Bi-variate correlation between items. Post-Session to the left. Post-Video to the right. Correlations above or equal to 0.5 are in bold.

		Post - Session											Post - Video										
		Mental	Physical	Temporal	Performance	Effort	Frustration	Demand	Supply	Understanding	Observing	Comprehension	Mental	Physical	Temporal	Performance	Effort	Frustration	Demand	Supply	Understanding	Observing	Comprehension
NASA TLX	Mental	-											-										
	Physical	.44	-										.45	-									
	Temporal	**.76**	.37	-									**.80**	.40	-								
	Performance	-.28	-.35	-.16	-								-.14	-.31	-.07	-							
	Effort-	**.87**	.40	**.79**	-.26	-							**.86**	.46	**.79**	-.22	-						
	Frustration	.49	.48	.39	**-.51**	**.51**	-						.36	.44	.35	**-.50**	.44	-					
SART	Demand	**.62**	.15	**.54**	-.13	**.59**	.29	-					**.60**	.20	**.58**	-.06	**.54**	.13	-				
	Supply	**.78**	.37	**.71**	-.19	**.75**	.38	**.76**	-				**.74**	.36	**.70**	-.13	**.71**	.27	**.77**	-			
	Understanding	-.34	-.31	-.20	.44	-.33	**-.55**	-.29	-.37	-			-.27	-.32	-.21	**.65**	-.32	**-.60**	-.24	-.33	-		
SA3	Observing	-.30	-.32	-.22	**.57**	-.30	**-.59**	-.20	-.27	**.70**	-		-.06	-.21	-.05	**.59**	-.12	-.45	-.07	-.11	**.68**	-	
	Comprehension	-.35	-.29	-.19	**.54**	-.31	**-.60**	-.24	-.29	**.73**	**.78**	-	-.25	-.28	-.17	**.64**	-.29	**-.57**	-.21	-.26	**.81**	**.75**	-
	Predict	-.24	-.18	-.14	.43	-.25	-.42	-.27	-.30	**.67**	**.64**	**.74**	-.25	-.30	-.22	**.59**	-.31	**-.54**	-.21	-.27	**.78**	**.69**	**.87**

Table 3
Correlation between post-session and post-video rating for each item.

	NASA TLX						SART			SA3		
	Mental	Physical	Temporal	Performance	Effort	Frustration	Demand	Supply	Understanding	Observing	Comprehension	Predict
r	.89	.89	.85	.56	.86	.77	.76	.79	.54	.43	.60	.59

Note: p < .001 for all correlations.

Table 4
Factor loadings resulting from operators' ratings of NASA TLX, SART, and SA3 items.

			Factor loadings			
			1	2	3	4
Post-Session	NASA TLX	Mental	**0.84**		0.30	
		Physical			**0.79**	
		Temporal	**0.79**			
		Performance			−0.31	−0.51
		Effort	**0.83**			
		Frustration	0.30		0.44	0.53
	SART	Demand	**0.78**			
		Supply	**0.85**			
		Understanding				**−0.74**
	SA3level	Observing				**−0.77**
		Comprehension		0.37		**−0.80**
		Predict		0.34		**−0.71**
Post-Video	NASA TLX	Mental	**0.86**			
		Physical			**0.75**	
		Temporal	**0.83**			
		Performance		0.61		−0.35
		Effort	**0.80**		0.30	
		Frustration		−0.45	0.43	0.40
	SART	Demand	**0.77**			
		Supply	**0.85**			
		Understanding		**0.80**		−0.34
	SA3Level	Observing		**0.75**		
		Comprehension		**0.90**		
		Predict		**0.82**		−0.30
	Variance explained (%)		.30	.17	.10	.15

Note: Factor loadings <0.3 are suppressed. Factor loadings above 0.70 is in bold.

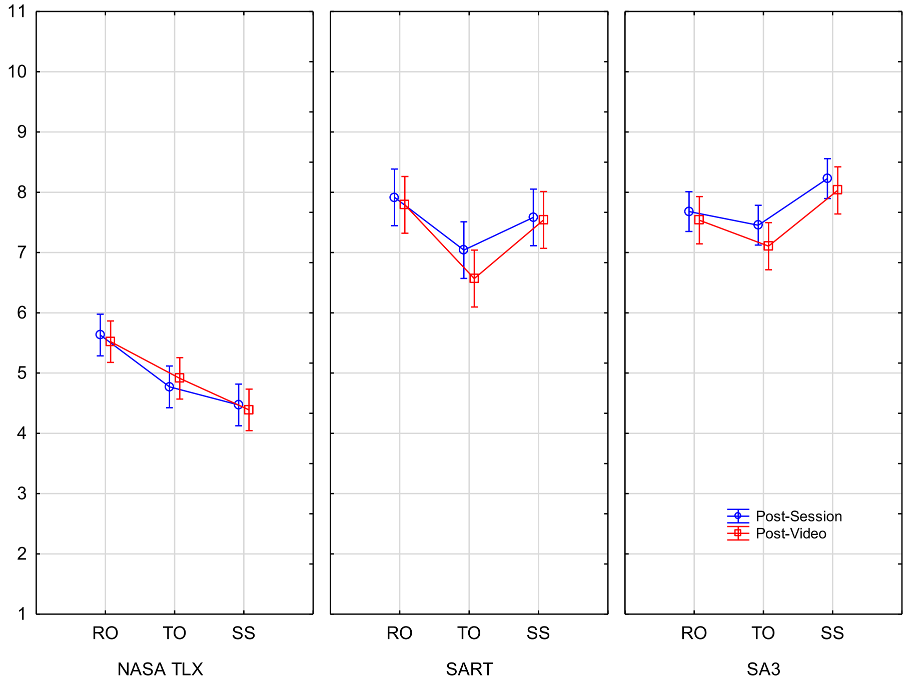

Fig. 3. *NASA TLX index, SART index and SA3 index, post-session, and post-video for each control room position. RO = Reactor Operator, TO = Turbine Operator, SS = Shift Supervisor. Mean and 95% confidence intervals.*

post-hoc test revealed that reactor operators' ratings significantly exceeded those of both the turbine operators and the supervisors, p = .006 and p > .001, respectively. The turbine operators' and supervisors' ratings were not statistically different. For the SART, the effect of position was significant, F(2,213) = 7.34, p > .001, while there was no significant difference in rating post-session versus post-video, F(1, 213) = 1.99, p = .16. Tukey's post-hoc test showed that both reactor operators' and supervisors' ratings exceeded turbine operators' rating, p < .001 and p = .02, respectively. The reactor operators' and supervisors' ratings were not statistically different. For SA3, the effect of position was significant F(2,213) = 6.91, p = .001, effect of post-session vs post-video F(1, 213) = 5.43, p = .02. Tukey's post-hoc test revealed that turbine operators rated lower than supervisors, p < .001. Reactor operators' ratings were not significantly different from turbine operators or supervisors.

Besides the overall analysis, the sensitivity to control room position (supervisor, reactor operator, turbine operator) of the post-session and the post-video ratings individually was analysed. Table 5 shows the F-statistic and the partial omega squared effect size resulting from a one-way ANOVA with position as the independent factor.

Generally, the results show that both post-session and post-video ratings were sensitive to varying loads of team positions. Particularly, the TLX physical demand, TLX frustration, and SART supply showed high sensitivity to the control room team position. Table 5 shows that the NASA TLX index sensitivity to team positions was quite similar for the post-session and the post-video ratings. Sensitivity of individual TLX items was slightly lower for the post-video than post-session, except for performance which showed slightly higher sensitivity for the post-video ratings. Of the individual items, not considering the indexes, the NASA TLX performance, SART understanding, and SA3 prediction showed higher sensitivity of the post-video ratings compared to the post-session ratings. Also, the SART understanding and the SA3 prediction were not significantly sensitive to the post-session rating but significantly sensitive to the post-video ratings.

3.6. Relating subjective ratings to performance

Validity evidence regarding the relationship with other criteria was investigated by relating the subjective ratings to the SCORE performance index. Table 6 presents the correlation between operator ratings, both post-session and post-video, and the SCORE performance index.

Table 6 shows that the workload items correlated negatively with task performance, while the subjective performance and situation awareness items correlated positively with task performance. The NASA TLX index and the majority of individual items, except performance, correlated negatively and significantly with the performance index. Also, the SART demand and supply correlations with task performance were negative. Due to relatively demanding scenarios, it was expected that an increase in experienced mental workload would relate negatively to performance. The TLX performance item was reversed to ease interpretation of the correlations—and the TLX performance correlated positively with the performance index. Also, the SA3 items and the SART understanding correlated positively with performance. Interestingly, only situation awareness items correlated significantly higher with performance in the post-video condition than in the post-session condition. Similar to situation awareness, the TLX performance correlation trended higher for the post-video rating than for the post-session, 0.28 and 0.38, respectively, but this difference was not significant.

Table 5
Sensitivity to control room team position. F- statistic and partial eta squared effect size. Post-session and post-video ratings.

		Post-session		Post-Video	
		F	Partial eta^2	F	Partial eta^2
NASA TLX	TLX Index	13.13***	.11	11.17***	.10
	Mental	7.25***	.06	4.82***	.04
	Physical	40.35***	.28	38.81***	.27
	Temporal	12.10***	.10	7.66***	.07
	Performance	10.78***	.09	14.64***	.12
	Effort	6.45***	.06	5.25***	.05
	Frustration	16.88***	.14	11.54***	.10
SART	SART Index	3.42***	.03	7.26***	.06
	Demand	7.74***	.07	5.32***	.05
	Supply	14.56***	.12	11.28***	.10
	Understanding	2.80***	.03	5.75***	.05
SA3	SA3 Index	5.72***	.05	5.41***	.05
	Observing	7.70***	.07	3.03***	.03
	Comprehension	7.18***	.06	6.42***	.06
	Predict	.1.51***	.01	4.90***	.04

Note: *) p < .05, **) p < .01, ***) p < .001.

Table 6
Pearson bi-variate correlation between operators' post-session and post-video subjective rating and performance index.

		Correlation between operator rating and SCORE performance index	
		Post-Session ratings	Post-Video ratings
NASA TLX	TLX Index	-.23**	-.29***
	Mental	-.17*	-.15*
	Physical	-.22**	-.20**
	Temporal	-.06	-.12
	Performance	.28***	.38***
	Effort	-.10	-.13
	Frustration	-.21**	-.31***
SART	SART Index**	.04	.30***
	Demand	-.01	-.10
	Supply	-.16*	-.21**
	Understanding**	.24***	.45***
SA3	SA3 Index**	.24***	.48***
	Observing***	.23**	.53***
	Comprehension*	.23**	.39***
	Predict*	.20**	.37***

Note: Significance of correlation, and significance of item difference between post-session and post-video correlation: *) p < .05, **) p < .01, ***), p < .001.

For relating workload and situation jointly to performance, multiple regressions of both post-session rating and post-video ratings were performed with the SCORE performance index as dependent variable. The resulting overall model fit and beta weights are presented in Table 7.

Table 7 shows that both the model containing the post-session ratings and the post-video ratings were statistically significant. The adjusted R^2 was 0.07 for the post-session ratings and 0.22 for the post-video ratings. Only the SA3 index was statistically significant, and the beta weight increased substantially in the post-video regression. The beta weight for the TXL index was in the expected direction but not statistically significant. Interestingly, the SART index behaved similar to the TLX workload index.

4. Discussion

This study investigated operators' subjective workload and situation awareness ratings just after the scenario was completed (post-session) and after operators' video-review of the scenario (post-video) to provide evidence for interpretations of this type of ratings. Based on the theoretical basis of the measures, it was hypothesised that operators' post-video ratings, having an external reference, would differ from post-session ratings, while items involving introspection would be similarly rated post-session and post-video. The results supported this hypothesis. Factor analysis resulted in a mental effort factor defined by both post-session and post-video ratings, while the analysis identified separate factors for operators' post-session perceptions of situation awareness and their post-video perceptions of situation awareness. The NASA TLX item frustration was related to other TLX dimensions and performance,

Table 7
Multiple regression of TLX index, SART, and SA3 on performance. Post-session and post-video ratings.

	Post-Session				Post Video			
	Adj R^2	F(3,212)	β	t(212)	Adj R^2	F(3,212)	β	t(212)
Overall Model	.07	6.64***			.22	21.77***		
TLX Index			-.07	−1.66			-.04	−1.17
SART Index			-.05	−1.65			-.02	−0.52
SA3 Index			.14	2.80**			.22	5.35***

as suggested by theory. However, the NASA TLX performance could not clearly be interpreted as an indication of operator self-regulation of effort. The SART items demand and supply were correlated with workload but not with situation awareness items, and the SART element of demand–supply was not related to SART understanding. The operators' rating of the scale based on Endsley's three-level theory of situation awareness was distinct from the operator's rating of workload and was substantially positively correlated with operator task performance.

4.1. The Nasa TLX

The study results did not support subjective ratings being able to capture the theoretical distinction between demand and effort. Similar findings have been reported in the literature (Hendy, 1995; Braarud, 2020). The factor analysis suggested one factor defined by mental demand, temporal demand, and effort regardless of operators' ratings being performed post-session or post-video. Although the distinction between demand and effort is theoretically sound (Gopher and Donchin, 1986; Hart and Staveland, 1988), it seems that the operator's rating of both mental demand and effort represents the mental effort invested. The NASA TLX scale description for mental demand (Hart and Staveland, 1988, Figure 8) reads "How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?". The first part of the description may guide the operator to introspect rather than looking outward at task demand; consequently, mental demand is assessed similarly to the operator's perception of mental effort. As hypothesised, plausibly due to the operator's assessment relying on introspection, the correlation between effort and operator task performance was not influenced by being informed about the actual scenario demand.

Complex motivational processes may be difficult to entangle from subjective scales alone, but the results of operator ratings did not support the interpretation of the NASA TLX performance as an indicator of self-regulation of effort. The TLX performance correlation with either effort or mental demand was low. However, performance substantially correlated with frustration (−0.53 and −0.52, post-session and post-video, respectively), suggesting that perception of poor performance was related to negative feelings. Similar to this study, the literature reports that the TLX performance item does not relate strongly to any of the other TLX items (Hendy, 1995; Bailey and Thompson, 2001; Braarud, 2020), and a hypothetical interpretation is that the TLX performance item represents how satisfied one is with one's own performance rather than self-regulation of workload. Adding to being correlated with performance, frustration correlated substantially with all other TLX items and loaded broadly on several factors—which can be interpreted as theoretically proposed, the psychological impact of workload.

A somewhat surprising result, given the highly mental characteristics of modern control room work, was the identification of a physical demand factor. Looking at the NASA TLX scale description for physical demand (Hart and Staveland, 1988, Figure 8), it reads "How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?". Given that the operators were seated at computerised workstations, it is hypothesised that they rated this item based on the physical human–computer interaction, such as navigating between process formats, navigating computerised procedures, and acknowledging alarms. This interpretation corresponds with findings relating human–computer interaction to operator workload (Lin et al., 2013; Braarud et al., 2020). Being informed about the scenario's task demand and reviewing one's own performance seemed to have no effect on the operator's rating of physical demand. This seems reasonable since the scenario replay likely did not result in any new insights about the scenario's physical demands. It is also interesting to note that the operators rated physical demand as relatively low, which corresponded to the mental characteristics of control room work. The item was also highly sensitive to the varying load of the team position, e.g., reflecting the physical human–machine interaction related to the reactor operators' relatively high information load.

4.2. SART

The results question the interpretation of the SART index, as an indicator of situation awareness. Similar to other studies (Hendy, 1995; Selcon et al., 1991; Loft et al., 2015; Braarud, 2020), this study found that the rating of demand and supply were related to the NASA TLX workload items rather than to situation awareness items. The rating of both demand and supply seems to involve introspection. Similar to TLX mental demand and effort, both the post-session and post-video ratings of the SART demand and supply loaded on the same factor. A factor that can be interpreted as mental effort. However, The SART understanding item loaded on the non-informed and informed situation awareness factors as expected, and understanding correlated with operator task performance.

The challenging part of the SART index seems to be the demand–supply element. There is a lack of evidence that the rating of demand–supply should modify situation awareness beyond the explicit rating of understanding. Unfortunately, the post-session demand–supply element of the SART formula did not relate to the understanding item or any of the SA3 situation awareness items. Also, informing the operators of the scenario demand (the post-video evaluation) did not influence the operators' rating of demand, such that the demand–supply element related significantly to the situation awareness ratings. The study's results on SART correspond with previous research (Endsley et al., 1998; Salmon et al., 2009), suggesting that the use of the SART index as a measure of situation awareness is questionable. The SART index may behave differently depending on the level of mental workload and may, in some cases, behave as a measure of workload rather than situation awareness (Pierce et al., 2008; Loft et al., 2015). Consequently, the extent to which the index represents situation awareness more accurately than the item understanding may not be clear. A reasonable interpretation of SART seems to be to interpret the dimensions individually—understanding indicating situation awareness, and demand and supply both indicating mental effort.

4.3. SA3

The SA3 rating of situation awareness was clearly distinct from workload. The SA3 items loaded highly on the two situation awareness factors and did not load on the workload factors. As hypothesised, being informed about scenario demand, and reviewing one's own performance influenced operators' rating of SA3 situation awareness—the non-informed ratings (post-session) and the informed rating (post-video) defined separate factors. The SA3 index was also more highly related to operator task performance than the SART index. Also, as expected, the informed rating (post-video) of SA3 was more closely related to operator task performance than the non-informed (post-session) rating. A probable explanation for this result is that, being informed, the operators could base their rating on what actually occurred in the scenario. As such, informed rating compared to non-informed rating would be hypothesised to be more closely related to a hypothetical true measure of situation awareness. Regarding the SA3 measures' three levels, the bi-variate correlations between the three items suggest that operators did not strongly discriminate these three items. The post-session bi-variate correlations ranged from 0.68 to 0.80, which is quite similar to correlations reported for similar items by Matthews et al. (2002), ranging from 0.66 to 0.74. Theoretically, the three levels should be related to each other (Matthews et al., 2002; Endsley et al., 2000). However, separate processes may be involved in the three levels, and human–system interface conditions may influence these three SA elements differently (Parasuraman et al., 2008; Endsley, 2000b). A plausible explanation for the relatively high bi-variate correlations is that the study was not explicitly designed to distinctively influence the three levels of situation awareness and that the overall rating of the relatively long scenarios made it difficult for the operators to separate the levels.

The positive correlation between SA and operator task performance increased significantly for post-video ratings compared to post-session ratings (post-session correlation ranged from 0.20 to 0.24, while post-video correlations ranged from 0.37 to 0.53). Also, the correlation between SA3 observing and operator task performance increased more from post-session to post-video rating than for the other two SA3 items. This indication of a somewhat distinct meaning of the SA3 observation plausible related to the scenario replay provided relatively concrete information on the operator's observation of malfunction symptoms, while the scenario replay provided an improved general basis for the rating of comprehension and prediction. However, the post-video correlations actualise the question of to what extent operator ratings represent perceptions of own performance rather than situation awareness (Endsley et al., 1998; Endsley, 2020). The increased correlation could mean that the informed rating approached objective situation awareness—which in supervisory control settings is assumed to be relatively highly related to task performance. Alternatively, the post-video SA rating included an element of an informed task performance rating. While subjective SA and performance seem distinguishable, future research needs to investigate whether informed subjective SA rating approaches objective SA or rather represents elements of operator task performance.

4.4. Sensitivity of subjective measures

Corresponding with the literature (O'Donnel and Eggemeier, 1986; Endsley et al., 1998; Hart, 2006), the results of this study suggest that subjective measures are sensitive. The NASA TLX index was more sensitive to the control room position than the SART index and the SA3 index. The higher sensitivity of the NASA TLX seems reasonable due to the scenarios being designed for different task loads on the control room positions. Both post-session and post-video ratings were sensitive to the control room position. However, ratings may be interpreted differently, although they have similar sensitivity. Looking at the SA3 index as an example, the effect size was 0.05 for both the post-session and post-video ratings. However, the SA3 post-session and post-video ratings loaded on different factors, and the post-video compared to the post-session SA3 index was more significantly correlated with the operator task performance index. This type of result reminds us that sensitivity does not equal validity, and this might be well worth emphasising regarding subjective ratings of complex phenomena.

4.5. Study limitations and future research

The study included a relatively modest sample of 18 operators from six control room teams. This relates to the practical challenge of recruiting professional control room operators for full-scope simulator studies, and future research should investigate the replicability of the results and investigate the extent to which the results generalise to less dynamic non-supervisory work. The operators' informed ratings (post-video) were performed after both being informed about scenario demand and after reviewing own performance. The study did not investigate to what extent either being informed or reviewing one's own performance influenced subjective ratings. However, the study demonstrated that being informed about scenario demand and reviewing one's own performance affected ratings of items involving the perception of external events but limited so for ratings involving introspection.

The SA3 situation awareness measure was developed for this study. The results and previous studies (McGuinness and Foy, 2000; Matthews and Beal, 2002) suggest that further investigation of this type of subjective situation awareness measure is desirable. Future studies are needed to investigate the degree to which SA3 is correlated with objective situation awareness measures, subjective performance, and confidence in one's own situation awareness to better assess its validity as an indicator of operator situation awareness (Endsley, 2020). Such studies could also investigate to what extent subjective ratings can distinguish between Endsley's (1995a) three levels of situation awareness. It can also be noted that the scale end points can preferably be reversed compared to the items used in this study. The study utilised the so-called 3D quick version of the SART measure. An application of the 10-item version (Taylor, 1990) might reveal nuances of the SART measures not captured in this study. Future studies could also investigate to what extent informed subjective ratings compared to non-informed subjective ratings are closer related to objective measures. Hence, future research could address whether more efficient approaches than the scenario replay applied in this study can adequately inform participants' subjective ratings.

It is also worth considering that overall subjective ratings, such as those investigated in this study, are generally not very good at capturing the detailed dynamics of work, nor do they accurately measure human factor phenomena of interest (Lysaght et al., 1989; Endsley et al., 1998). However, establishing the adequate interpretation of subjective ratings, e.g., what are the measures valid for, is important in guiding the selection of measure and interpretation of their results. To assess complex mental work, subjective measures can serve a purpose in combination with other types of measures (Lysaght et al., 1989; Annett, 2002) and can be applied to screening scenarios or performance episodes for further analysis (Meister, 1976).

5. Conclusion

The study found that subjective assessment involving introspection seems to have a robust interpretation across conditions, while the interpretation of items involving perception of external events depends on the participants being informed about what actually happened in the scenario. To obtain valid subjective measures of situation awareness and performance it is recommended to inform participants of system malfunctions implemented in the scenario and system performance implications of their actions. Operators' ratings seem to entangle separated and interpretable constructs for workload, situation awareness, and performance. However subjective workload ratings did not distinguish between mental demand and effort. There is evidence of the multidimensionality of the NASA TLX measure applied to complex cognitive work. Operators' ratings could distinguish between mental effort, physical activity, subjective performance, and frustration. The results of the study question the SART index as a measure of situation awareness.

The results did not support an interpretation of the SART index's demand–supply element as an indicator of situation awareness. However, the interpretation of the SART dimensions individually seems reasonable—demand and supply indicate operator effort, while the dimension understanding represents situation awareness. A subjective measure based on Endsley's (1995a) three-level situation awareness theory showed promising results, adding to previous studies of similar subjective measures (McGuinness and Foy, 2000; Matthews and Beal, 2002). The measure was distinct from workload and related to operator task performance. The promising results warrant future research to determine the validity and utility of this type of subjective measure. Future research could investigate whether informing participants about what actually occurred in a scenario results in subjective ratings of situation awareness and performance that approach the results of objective measures. To what extent and how to inform participants adequately and efficiently in subjective ratings could also be further researched. Finally, future research could investigate whether the study's findings can be replicated in related domains and if the findings extend to less dynamic non-supervisory work.

CRediT authorship contribution statement

Per Øivind Braarud: Conceptualization, Methodology, Formal analysis, Writing - original draft, Writing - review & editing.

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.

Acknowledgements

The data utilised in this study was collected for work funded by the OECD Halden Reactor Project, Norway.

References

Annett, J., 2002. Subjective rating scales: science or art? Ergonomics 45 (14), 966–987.

Bailey, L., Thompson, R., 2001. The TLX: one or more constructs. In: Proceedings of the 11th International Symposium of Aviation Psychology. The Ohio State University, Columbus.

Braarud, P.Ø., 2020. An efficient screening technique for acceptable mental workload based on the NASA Task Load Index—development and application to control room validation. Int. J. Ind. Ergon. 76 https://doi.org/10.1016/j.ergon.2019.102904.

Braarud, P.Ø., Kirwan, B., 2011. Task complexity: what challenges the crew and how do they cope? In: Skjerve, A.B., Bye, A. (Eds.), Simulator-based Human Factors Studies across 25 Years: the History of the Halden Man-Machine Laboratory. Springer-Verlag, pp. 233–251.

Braarud, P.Ø., Svengren, H., 2020. Considerations for the Design of Human Factors Validation Scenarios. HWR-1300. Halden, Norway. Organisation for Economic Co-operation and Development (OECD), Halden Reactor Project.

Braarud, P.Ø., Eitrheim, M.H.R., Fernandes, A., 2015. "SCORE" - an integrated performance measure for control room validation. Proceedings of the Ninth American Nuclear Society International Topical on Nuclear Plant Instrumentation, Controls and Human-Machine Interface Technologies (NPIC & HMIT 2015). American Nuclear Society, La Grange, IL.

Braarud, P.Ø., Eitrheim, M.H.R., Holmgren, L., McDonald, R., 2016. Review of the SCORE Measure for the Assessment of Safe Control Room Operation: A HAMMLAB Study of the Content Validity and Usability. HWR-1175. Halden, Norway. OECD Halden Reactor Project.

Braarud, P.Ø., Bodal, T., Hulsund, J.E., Louka, M.N., Nihlwing, C., Nystad, E., Svengren, H., Wingstedt, E., 2020. An investigation of speech features, plant system Alarms, and operator–system interaction for the classification of operator cognitive workload during dynamic work. Hum. Factors. https://doi.org/10.1177/0018720820961730.

Byers, J.C., Bittner Jr., A.C., Hill, S.G., 1989. Traditional and raw task load index (TLX) correlations: are paired comparisons necessary?. In: Advances in Industrial Ergonomics and Safety. Taylor & Francis, London, pp. 481–485.

Charles, R.L., Nixon, J., 2019. Measuring mental workload using physiological measures: a systematic review. Appl. Ergon. 74, 221–232. https://doi.org/10.1016/j.apergo.2018.08.028.

Colle, H.A., Reid, G.B., 2005. Estimating a mental workload redline in a simulated air-to-ground combat mission. Int. J. Aviat. Psychol. 15 (4), 303–319. https://doi.org/10.1207/s15327108ijap1504_1.

De Winter, J., 2014. Controversy in human factors constructs and the explosive use of the NASA-TLX: a measurement perspective. Cognit. Technol. Work 16, 289–297. https://doi.org/10.1007/s10111-014-0275-1.

Endsley, M.R., 1995a. Towards a theory of situation awareness in dynamic systems. Hum. Factors 37, 32–64.

Endsley, M.R., 1995b. Measurement of situation awareness in dynamic systems. Hum. Factors 37, 65–84.

Endsley, M.R., 2000a. Theoretical underpinnings of situation awareness: a critical review. In: Endsley, M.R., Garland, D.J. (Eds.), Situation Awareness Analysis and Measurement. LEA, Mahwah, NJ, pp. 3–32.

Endsley, M.R., 2000b. Direct measurement of situation awareness: validity and use of SAGAT. In: Endsley, M.R., Garland, D.J. (Eds.), Situation Awareness Analysis and Measurement. Lawrence Erlbaum Assoc, Mahwah, pp. 147–173.

Endsley, M.R., 2020. The divergence of objective and subjective situation awareness: a meta-analysis. Journal of Cognitive Engineering and Decision Making 14, 34–53. https://doi.org/10.1177/1555343419874248.

Endsley, M.R., Selcon, S.J., Hardiman, T.D., Croft, D.G., 1998. A comparative evaluation of SAGAT and SART for evaluations of situation awareness. In: *Proceedings Of the Human Factors And Ergonomics Society 42nd Annual Meeting*. Human Factors and Ergonomics Society, Santa Monica, CA, pp. 82–86.

Endsley, M.R., Holder, L.D., Leibrecht, B.C., Garland, D.J., Wampler, R.L., Matthews, M.D., 2000. Modeling and Measuring Situation Awareness in an Infantry Operational Environment (Report 1 753). U.S. Army Research Institute for the Behavioral Sciences, Alexandria, VA.

Fracker, M.L., 1991. Measures of Situation Awareness: Review and Future Directions (Report No. AL-TR-1991-0128). Armstrong Laboratories, OH. Wright-Patterson Air Force Base.

Gan, Y., Dong, X., Zhang, Y., Zhang, X., Jia, M., Liu, Z., Li, Z., 2020. Workload measurement using physiological and activity measures for validation test: a case study for the main control room of a nuclear power plant. Int. J. Ind. Ergon. https://doi.org/10.1016/j.ergon.2020.102974.

Gopher, D., Donchin, E., 1986. Workload - an examination of the concept. In: Boff, K.R., Kaufman, L., Thomas, J.P. (Eds.), Handbook of Perception and Human Performance, Vol II, Cognitive Processes and Performance. Wiley & Sons, New York.

Grier, R.A., 2015. How high is high? A meta-analysis of NASA-TLX global workload scores. Proceedings of the Human Factors and Ergonomics Society 59th Annual Meeting 32, 150–154.

Hart, S., 2006. Nasa-task load index (Nasa-TLX); 20 years later. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 50 https://doi.org/10.1177/154193120605000909.

Hart, S.G., Staveland, L., 1988. Development of the NASA task load index (TLX): results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (Eds.), Human Mental Workload. North-Holland, Amsterdam, pp. 139–183.

Hendy, K.C., 1995. Situation awareness and workload: birds of a feather*? Situation awareness: Limitations and Enhancement in the aviation environment*. Neuilly sur Seinne, FR: AGARD 21–1, 22–27.

IFE, 2021. HAMMLAB. Institute for Energy Technology. https://ife.no/en/laboratory/hammlab/.

Kim, J., Mueller, C.W., 1978. Factor Analysis. Statistical methods and practical issues. In: Sage University Papers on Quantitative Applications in the Social Sciences. Sage Publications, Beverly Hills and London, 07-014.

Kirk, R.E., 1995. Experimental Design: Procedures for the Behavioral Sciences. Brooks/Cole Publishing Company.

Lin, T., Li, X., Wu, Z., Tang, N., 2013. Automatic cognitive load classification using high-frequency interaction events. Int. J. Technol. Hum. Interact. 9, 73–88. https://doi.org/10.4018/jthi.2013070106.

Loft, S., Bowden, V., Braithwaite, J., Morrell, D.B., Huf, S., Durso, F.T., 2015. Situation awareness measures for simulated submarine track management. Hum. Factors 57, 298–310.

Lysaght, R.J., Hill, S.G., Dick, A.O., Plamondon, B.D., Linton, P.M., Wierwille, W.W., Zaklad, A.L., Bittner Jr., A.C., Wherry Jr., R.J., 1989. Operator Workload: Comprehensive Review and Evaluation of Operator Workload Methodologies (ARI Technical Report 851). United States Army Research Institute, Fort Bliss, TX, USA.

Matthews, M.D., Beal, S.A., 2002. A field test of two methods for assessing infantry situation awareness. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Sage, Santa Monica, CA, pp. 352–356.

Matthews, G., Winter, J.D., Hancock, P.A., 2020. What do subjective workload scales really measure? Operational and representational solutions to divergence of workload measures. Theor. Issues Ergon. Sci. 21 (4), 369–396.

McGuinness, B., 2021. The Crew Awareness Rating Scale (CARS). Unpublished manuscript.

McGuinness, B., Foy, L., 2000. A subjective measure of SA: the crew awareness rating scale (CARS). In: Proc. Of the First Human Performance, Situation Awareness, and Automation Conference. Savannah, Georgia.

Meister, D., 1976. Behavioral Foundations of System Development. John Wiley &Sons, New York.

Messick, S., 1990. Validity of Test Interpretation and Use. Research Report 90-11. Education Testing Service, Princeton, New Jersey.

Messick, S., 1995. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am. Psychol. 50 (9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741.

Muckler, F.A., Seven, S.A., 1992. Selecting performance measures. Objective' versus 'subjective' measurements Human Factors 34, 441–455.

Murphy, K.R., Davidshofer, C.O., 1994. Psychological Testing: Principles and Applications, third ed. Prentice-Hall, Englwood Cliffs, NJ.

Nygren, T.E., 1991. Psychometric properties of subjective workload measurement techniques: implications for their use in the assessment of perceived mental workload. Hum. Factors 33 (1), 17–33.

Øwre, F., 2011. The history of HAMMLAB – 25 years of simulator-based studies. In: Skjerve, A.B., Bye, A. (Eds.), Simulator-based Human Factors Studies across 25 Years: the History of the Halden Man-Machine Laboratory. Springer-Verlag, pp. 13–42.

O'Donnell, R.D., Eggemeier, F.T., 1986. Workload assessment methodology. In: Boff, K. R., Kaufman, L., Thomas, J.P. (Eds.), Handbook of Perception and Human Performance, Vol II. Wiley & Sons, pp. 1–49.

O'Hara, J.M., Higgins, J.C., Fleger, S.A., Pieringer, P.A., 2012. Human Factor Engineering Program Review Model. NUREG-0711, Rev.3. U.S. Nuclear Regulatory Commission, Washington DC, USA.

Parasuraman, R., Sheridan, T.B., Wickens, C.D., 2008. Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. Journal of Cognitive Engineering and Decision Making 2 (2), 140–160.

Pierce, R.S., Strybel, T.Z., Vu, K.P.L., 2008. Comparing situation awareness measurement techniques in a low fidelity air traffic control simulation. In: Proceedings of the 26th International Congress of the Aeronautical Sciences (ICAS). Anchorage, AS.

Reid, G.B., Colle, H.A., 1988. Critical SWAT values for predicting operator overload. In: Proceedings of the Human Factors Society 32nd Annual Meeting. Human Factors Society, Santa Monica, CA, pp. 1414–1418.

Reid, G.B., Nygren, T.E., 1988. The Subjective Workload Assessment Technique: a scaling procedure for measuring mental workload. In: Hancock, P.A., Meshkati, N. (Eds.), Advances in Psychology, 52. Human Mental Workload. North-Holland, Oxford, England, pp. 185–218.

Salmon, P.M., Stanton, N.A., Walker, G.H., Jenkins, D., Ladva, D., Rafferty, L., 2009. Measuring Situation Awareness in complex systems: comparison of measures study. Int. J. Ind. Ergon. 39 (3), 490–500.

Selcon, S.J., Taylor, R.M., Koritsas, E., 1991. Workload or situational awareness?: TLX vs. SART for aerospace systems design evaluation. In: Proceedings of the Human Factors Society's 35th Annual Meeting. Human Factors Society, Santa Monica, CA, USA, pp. 62–66.

Taylor, R.M., 1990. Situation awareness rating technique (SART): the development of a tool for aircrew systems design. In: A GARD-CP-478, Situation Awareness in Aerospace Operations. Advisory Group for Aerospace Research & Development, Neuilly Sur Seine, France, 3-1 to 3-17.

Taylor, R.M., Selcon, S.J., 1994. Situation in mind: theory, application and measurement of situational awareness. In: Gilson, R.D., Garland, D.J., Koonce, J.M. (Eds.), Situational Awareness in Complex Settings. Embry-Riddle Aeronautical University Press, Daytona Beach, FL, pp. 69–78.

Vidulich, M.A., 2000. The relationship between mental workload and situation awareness. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 460–463.

Vidulich, M.A., Tsang, P.S., 2015. The confluence of situation awareness and mental workload for adaptable human–machine systems. Journal of Cognitive Engineering and Decision Making 9 (1), 95–97. https://doi.org/10.1177/1555343414554805.

Yeh, Y.Y., Wickens, C.D., 1988. Dissociation of performance and subjective measures of workload. Hum. Factors 30 (1), 111–120.

Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A., 2015. State of science: mental workload in ergonomics. Ergonomics 58 (1), 1–17. https://doi.org/10.1080/00140139.2014.956151.