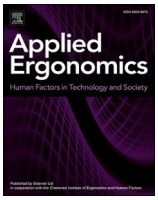
Applied Ergonomics 97 (2021) 103500

Contents lists available at ScienceDirect

Applied Ergonomics

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



Comparing control room operators' and experts' assessment of team performance using structured task-specific observation protocols and scenario replay

Per Øivind Braarud

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

ARTICLE INFO

Keywords:
Self-assessment
Expert assessment
Assessment method
Team performance

ABSTRACT

Operators' self-assessment has received limited interest within process control or human-system evaluation. Research on self-assessment has been criticised for poor assessment methodology, and consequently, its status is unclear. This study hypothesised that, given adequate assessment methods (such as task-specific assessment items and scenario replay), we could observe relatively accurate self-assessment results. Eighteen licensed operators and two experts assessed team performance in six nuclear control room scenarios. The results reveal an overall agreement between operators and experts, measured by the intraclass correlation coefficient, ranging from 0.60 to 0.70, which lies close to the intraclass correlation coefficient of 0.75 for the experts. This demonstrates potential for achievement of relatively accurate operator self-assessment for complex work. The agreement varied in a similar manner for both expert agreement and operator-expert agreement across eight performance dimensions. In addition, the operators' self-assessment provided additional information beyond observer assessment in identifying non-acceptable performance items.

1. Introduction

Operators' performance and work conditions receive considerable interest due to their importance in ensuring safe and efficient operation. However, team performance assessment in complex human-machine settings are challenging to undertake (Vreuls and Obermayer, 1985; Rosen et al., 2008) and involve substantial degrees of judgement regarding the safety implications of observed behaviours (Hall and Brannick, 2009). Performance dimensions addressed in operator training and human-system evaluation of nuclear control rooms include observation of plant status, situation assessment, strategy and response planning, implementation of control actions, and teamwork (O'Hara et al., 2012; Xu et al., 2018; Simonsen and Osvalder, 2018). Frequently, subject matter experts observe team performance and use structured observation protocols developed via task analysis in their assessment of performance (Landy and Farr, 1983; Hall and Brannick, 2009; Wildman et al., 2013). However, operators' self-assessment might provide valuable perspectives to the information gained from observer assessment (Muckler and Seven, 1992; Sinclair, 1995; Annett, 2002) and serve a practical purpose due to the limited availability of expert resources (Mete and Brannick, 2017; Wieck et al., 2018). Furthermore, self-assessment is an essential component of self-regulated learning (Gordon, 1991; Eva and Regehr, 2005; Van Loon, 2018).

Operators' self-assessment has been the subject of limited research within process control, whereas self-assessment regarding team performance has been extensively researched in the medical domain (Weller et al., 2013; Marriage and Kinnear, 2016; Ganni et al., 2017) and also in aviation (Gontar and Hoermann, 2015). Generally, self-assessment is conducted through comparing self-observed performance against a certain standard (Colthart et al., 2008; Van Loon, 2018). Clearly, one must assume the validity of self-assessment when applying its use (Boud, 1995; Van Loon, 2018).

Benefits of self-assessment include improved motivation for competence development and improved commitment to performance standards (Marienau, 1999; Gordon, 1992). Therefore, self-assessment is also a form of quality assurance supporting safe and efficient performance (Arora et al., 2011). In the area of workload, Muckler and Seven (1992) suggest that an advantage of subjective workload assessment techniques is that the operator can be aware of increased effort as related to potentially negative performance effects. Such an advantage could also be relevant for performance assessment based on the operator's personal experience of cognitive challenges. Another practical motivation for investigating self-assessment is the limited availability of experts and the resources needed to utilise expert assessment (Muckler

E-mail addresses: Per.Oivind.Braarud@ife.no, per.oivind.braarud@hrp.no.

https://doi.org/10.1016/j.apergo.2021.103500
Received 9 April 2020; Received in revised form 7 June 2021; Accepted 8 June 2021
Available online 5 July 2021

and Seven, 1992; Annett, 2002; Mete and Brannick, 2017). The limited number of assessors also challenges the reliability of measurement in terms of inter-rater agreement (Vreuls and Obermayer, 1985; Sinclair, 1995; Hall and Brannick, 2009).

When compared to observer assessment, self-assessment is generally credited with low or moderate accuracy (Eva and Regehr, 2005; Davis et al., 2006; Van Loon, 2018). Participants tend to overestimate their own performance when compared to expert observers' assessment (Davis et al., 2006; Wieck et al., 2018; Nayar et al., 2020). It has also been noticed that while poor performers overestimate their performance, top performers underestimate theirs (Kruger and Dunning, 1999; Ehrlinger et al., 2008; Boud et al., 2015). It is not clear to what extent these biases apply to professional control room operators' self-assessment. However, self-assessment accuracy increases with increased work experience (Nayar et al., 2020) and with increased competence (Kruger and Dunning, 1999; Ehrlinger et al., 2008). Additionally, self-assessment accuracy varies according to the performance dimension assessed. Studies in the medical domain find moderate agreement between self-assessment and expert assessment of technical performance but no or limited agreement on the assessment of non-technical performance (Moorthy et al., 2006; Arora et al., 2011; Wieck et al., 2018). Similar results have been reported in aviation. Gontar and Hoerman (2014) found higher agreement for cognitive technical performance than for non-technical performance in this domain.

Explanations for poor or moderate accuracy of self-assessment include the following: limited meta-cognitive insight (Kruger and Dunning, 1999; Kim, 2018; Van Loon, 2018), lack of training on self-assessment procedures and techniques (Van Loon, 2018), and method issues, such as poor design of self-assessment tasks and standards (Ward et al., 2002; Colthart et al., 2008; Van Loon, 2018). With regard to method, several considerations frequently related to expert observers might be relevant whether the assessor is an expert or an operator self-assessing their performance. For simulator-based assessment, relevant factors include the amount and accuracy of information provided to the assessors, to what extent assessors attend to the same behaviour, and to what extent assessors make the same interpretations and inferences from the behaviours observed (Weber et al., 2013).

Recommended practices for performance assessment include the following: using structured, task-specific observation protocols to support the assessor's attention to relevant performance aspects (Rosen et al., 2008), providing adequate conditions for observation, whether online or offline (Hall and Brannick, 2009), training observers (Holt et al., 2002; Rosen et al., 2008), and involving subject matter experts in the development of such protocols (Hall and Brannick, 2009; Braarud et al., 2015). Regarding conditions for observation, Hall and Brannick (2009) suggest that assessors are more likely to provide reliable assessment when video recordings are used rather than live observation of actual work because video recordings can be paused and replayed (Wildman et al., 2013). These recommended practices have also been related to self-assessment. For example, specific self-assessment measures based on task analysis have been found to be more accurate than generic assessment because they better represent the performance domain and standards (Gaba et al., 1998; Brannick et al., 2002; Van Loon, 2018). Similarly, the advantage of using video has also been identified specifically for self-assessment (Gordon, 1991; Steinemann et al., 2012; Nayar et al., 2020).

Current literature provides limited insights into the utility and accuracy of professional control room operators' self-assessment regarding their performance. This study intended to investigate participating operators' self-assessment by comparing their assessment to those of expert observers. The study hypothesised that we could observe a comparable level of agreement between professional operators' self-assessment and assessment by experts. By using recommended assessment procedures, the study allowed operators to perform self-assessment in conditions similar to those frequently used for expert observers. After completing simulated scenarios, participating operators used task-specific observation protocol and a video/simulator log replay of the scenario to assess their own team's performance. Experts analysed team performance using the same assessment procedure.

2. Method

The relationship between self-assessment and observer assessment was studied in a simulator experiment at IFE's Halden Human-Machine Laboratory (Institute for Energy Technology [IFE], 2020). Both participating operators and expert observers used the recorded performance of the control room team for post-scenario assessment by applying the same performance assessment technique.

2.1. Participants

Eighteen licensed, male, control room operators participated in the study. They composed six control room teams, each consisting of a supervisor, a reactor operator, and a turbine operator. Five of the six teams consisted of colleagues working as a team at their home plant, while one of the teams was configured with operators from different home plant teams. The mean age was 42.7 years (SD = 11.06) for supervisors, 44.7 years (SD = 13.6) for reactor operators, and 30.3 years (SD = 5.65) for turbine operators. Their mean experience of control-room work was 17.2 years for supervisors (range 6–34); 12.67 years (range 7–35) for reactor operators, and 2.0 years (range 1–5) for turbine operators. 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 procedures for human participant protection.

The participants' technical competence covers two main parts of the nuclear power plant: the reactor side and the turbine side. The participants' education and training at their home plant is organised in such a way that a control room operator can progress from the position of turbine operator to a reactor operator, and then to the position of supervisor. Thus, both supervisors and reactor operators have detailed technical competences on the functioning of both the plant's turbine and reactor. While the turbine operators' overall technical competence covers the whole plant, including the reactor, their main competence is concentrated on the turbine. The supervisors had competence on team leadership, and all participants had competence on teamwork and communication procedures for control-room work. The participants regularly performed simulator-based training on both technical and non-technical skills at their home plant.

Two expert observers participated in the study. Both experts were former licensed control-room supervisors and training instructors from a plant relevant to the research simulator and to the participants' home plant. The experts' mean experience of control-room work was 15 years and their mean experience as subject-matter experts, 19 years. Both experts had extensive experience with performance-based evaluation, including the development of scenarios and specification of measures for human-factor validation and simulator research.

2.2. Performance assessment and scenario replay tool

The simulator sessions were recorded with the laboratory's video, audio, and data analysing tool for use in post-scenario assessment. The tool provided synchronised play of simulator logs, video and audio from a scenario completed in the simulator. The recording included the simulated plant's process development (alarms, process parameters, process events), operator process commands, navigation and interfaces accessed by the operator, videos of each operator workstation, as well as an overview video of the control room, and separate audio recordings from each of the control room operators. An assessor could play, pause, rewind, and forward the scenario during the performance assessment. Video recordings and interfaces addressed by the operators could be placed in separate windows and the volume of each audio could be

separately controlled.

The performance assessment used the Supervisory Control and Resilience Evaluation (SCORE) measure (Braarud et al., 2015; Braarud and Berntsson, 2016). SCORE is a framework for developing task-specific performance evaluation. The framework included a description of how to structure a scenario into events (performance episodes), a description of performance dimensions, a rating scale for assessing acceptability of performance, formatting for assessment sheets, and examples of concrete measures that have been applied to simulator sessions for nuclear control-room teams. The framework described two types of items: assessment items and observation items. Assessment items, the focus of the measure, considered the quality of team performance and were rated on a six-point scale. Observation items were included to support the assessors' overview of an event, as a basis for their assessment. Observation items considered whether the team detected detailed information or performed detailed actions. The measure has been evaluated by subject-matter experts (Braarud et al., 2016), applied in simulator experiments (Eitrheim et al., 2018) and used for human-factor validation in the nuclear industry (Gunnarsson et al., 2014; Braarud et al., 2019; Braarud, 2020). This study utilized the framework's performance dimensions: monitoring, interpretation, strategy, action, verification, teamwork, and goals. A second topic of the study regarding computerised plant overview (Braarud and Svengren, 2020) alerted this research to that fact that certain operator behaviour identified through the experiment preparations did not fully fit one of the previously described dimensions; this study therefore added the ad-hoc dimension "work process". The dimension was based on the concept of workers "finishing the design" (Rasmussen, 1986; Vicente, 1999) - operators adapt their way of utilizing the human-machine interface to fit the demands of the situation and to fit their situated goals.

Table 1
Operational definition of performance dimensions and example assessment items.

Dimension	Operational definition	Example assessment item
Monitoring	Observing the plant process and detecting deviating objects and systems, monitoring key process parameters	- Detects busbar X has been de-energised. - Detects condenser pressure increasing.
Interpretation	Making sense of the situation, diagnosing faults, looking ahead and foreseeing the plant development.	- Understands that B channel is incorrect by comparing its indications to the other three channels.
Strategy	Choosing an event procedure or deciding on a procedure path. Adapting the procedure guidance to the situation.	- Decides manual stop of the reactor. - Checks and considers the limitations from the technical specifications for plant system X.
Action	Manipulating plant's objects, system and functions. Action can also occur by ordering plant staff to perform them.	- Opens depressurisation valve X. - Orders field operator to adjust the cooling flow.
Verification	Verifying and checking the response from actions performed.	- Checks that the safety objects start following the start order. - Verifies that the emergency diesel generator starts.
Teamwork	Informing team members of the main plant process development observed, as well as important actions or expected plant progression.	- Communicates that reactor coolant train X has been actuated. - Alerts that the pressure in the condenser rises.
Goal	Overall goals of controlling the plant in terms of plant safety and production.	- Performs change of feedwater pumps such that down-regulation is avoided. - Identifies the safety systems failures and controls the reactor.
Work Process	Way of interacting with the human-machine interface. How the information is presented, and the controls are used and managed.	- Performance of alarm management. - Management of the display system for surveillance of safety technical specifications.

Table 1 shows simplified and summarised operational definitions of performance dimensions used when developing task-specific assessment sheets. The table includes examples of assessment items.

During the experiment preparations, task-specific assessment sheets for the scenarios were developed by one of the participating subject-matter experts. By running the scenarios in the simulator, using the operating procedures, the conduct of operation and the technical specifications provided a basis for identifying and specifying the SCORE items. When necessary, one of the experts consulted the other expert for a second opinion or verification of the proposed items. For the six scenarios (see below) the measure specification resulted in a total of 245 assessment items and 89 observation items. Table 2 shows the number of items identified per performance dimension for each of the scenarios.

The items specified were implemented in a computerised tool for application in the study. Fig. 1 illustrates the structure of the SCORE measure as implemented in the assessment tool for one event in a scenario.

Based on the SCORE documentation, the scale was defined as 1 = Strongly not acceptable, 2 = Not acceptable, 3 = Acceptability disputable, but probably *not* acceptable, 4 = Acceptability disputable, but probably acceptable, 5 = Acceptable, and 6 = Strongly acceptable. Acceptable was defined as the observed performance meeting the assessor's standards for control-room work. An observation item was scored simply as performed or not performed. This binary judgement was labelled "Yes" or "No". For all items, an assessor could judge an item as "Not Applicable" (NA). For example, an item could be not applicable due to a teams' strategy for handling an event, such as by down-prioritising and postponing a task until the scenario was ended. Finally, an assessor could state an uncertain basis for the assessment by ticking a question mark.

2.3. Performance assessment procedure

The assessment procedures and the tool were demonstrated and explained to the participating operators before the assessment. The explanation included an oral explanation of the SCORE method and walking through a short example scenario while demonstrating the scenario replay and the assessment software. Experimental staff were, upon request from the participants, available for assistance on the technical aspects of the assessment procedure, the assessment tool, or the scenario replay tool during the self-assessment.

The participating operators performed the self-assessment individually after each scenario. Immediately after completing the scenario and before the assessment, each team member received a description of the scenario. The scenario description included a brief explanation of the plant failures implemented in the scenario just performed. The team supervisor was instructed to talk through the scenario with the team to assure their understanding of the failures implemented in the scenario. This procedure was performed to provide the participants with an understanding of the scenario similar to what would be expected for a prepared expert. Fig. 2 shows a team member performing the self-assessment using the assessment tools. Each team member conducted self-assessment at a separate assessment station. The two experts performed the assessment of each scenario post-experiment separately, using the same scenario replay tool, and the same performance tool and assessment procedures as the participating operators.

2.4. Control-room simulator and scenarios

The nuclear power plant simulator at IFE's Halden Human Machine Laboratory (HAMMLAB) (IFE, 2020) is an advanced nuclear boiling water reactor. The HAMMLAB control room was equipped with digital interfaces and controls, and computerised operating procedures. The plant process simulation and the human-machine interfaces were comparable to a full-scope training simulator. Fig. 3 shows the control room

Table 2
Number of SCORE Assessment and Observation items per performance dimension and per scenario.

	Assessment items							Observation items						
	Sc 1	Sc 2	Sc 3	Sc 4	Sc 5	Sc 6	Sum	Sc 1	Sc 2	Sc 3	Sc 4	Sc 5	Sc 6	Sum
Monitoring	12	12	18	20	8	6	76	7	8	11	14	11	12	63
Interpretation		4	2	8	9	2	25							
Strategy	1	2	3		2	1	9		1			3	1	5
Action	2	3	4	3	3	5	20	2	3	1	4	2	4	16
Verification	5	5	4	4	4	9	31	2				1	2	5
Teamwork	1	8	6	8	7	12	42							
Goal	2	4	3	6	1	2	18							
Work Proc.	5	5	3	4	2	5	24							
Total	28	43	43	53	36	42	245	11	12	12	18	17	19	89

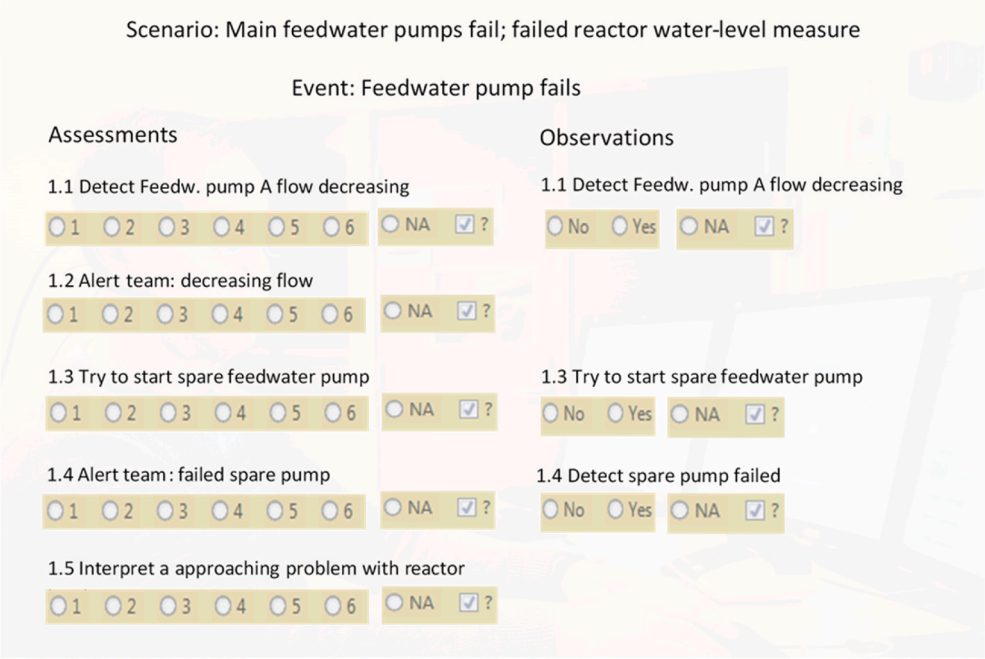


Fig. 1. SCORE items of assessment tool; one example event from one scenario.

Fig. 2. Participating operator performs self-assessment. Upper large screen provides scenario replay; lower screen provides assessment tool.

as configured for the study.

The simulation and test scenarios were controlled from an experimenter's gallery adjacent to the control room. Plant staff external to the control room, such as field operators, technical support and plant management, were role-played by a subject-matter expert. The orders of the control-room team to field operators out in the plant were

Fig. 3. Control room of HAMMLAB full-scale research simulator.

implemented in the simulator by the subject-matter expert during the scenario. The control-room team communicated with external staff by telephone, as they would do at their home plant.

Each team participated in six scenarios. The scenarios were developed by one of the participating subject-matter experts and discussed with the human-factors leader of the study (the author). Similar, but not identical, scenarios have been used in simulator research (Laumann et al., 2006; Eitrheim et al., 2018) and human-factor validation for the nuclear industry (Gunnarsson et al., 2014; Braarud, 2020). Each

scenario lasted from 20 to 45 min. The general scenario structure consisted of initial work with normal operation or periodic testing, including minor plant system failures, additional failures and aggravation of plant challenges leading to actuation of plant safety functions (reactor protection systems). The situation included several safety systems malfunctions. The control-room team would perform procedures to gain an overview of the plant's safety status and to develop a basis for choosing a strategy to mitigate the situation and prevent further degradation of safety systems. Some of the malfunctions were relatively unfamiliar to the teams. The set of scenarios included events frequently modelled in the safety analysis of nuclear power plants, such as loss of reactor coolant accidents (LOCA), loss of offsite power (LOOP), loss of turbine condenser, loss of main feedwater, and reactor containment isolation. As an example, Table 3 provides a brief overview of the events of two scenarios.

2.5. Interview

Each control-room team participated in a brief, semi-structured group interview of about 30 min. The interview was based on guiding questions about the self-assessment, as follows: "What are your experiences and your thoughts regarding the self-assessment?", "What do you think about the relevance of the SCORE items?", "Do you think this type of self-assessment is relevant for your training in the simulator?", and "How relevant were the scenarios for your competence development and for your work?".

2.6. Study procedure

Each control room team participated in the study for three consecutive days. The first half-day included an experiment briefing and training on the HAMMLAB computerised interfaces. After completing all scenarios, the interview was performed, and a debriefing according to the human participant protection procedures concluded their participation.

The participating teams were instructed to use work routines, operating and administrative procedures as they ordinarily would in their simulator training and competence assessment at their home plant, including routines for communication, teamwork and interacting with plant staff external to the control room. As described above, during the scenarios, external staff and their activities were role-played by a subject-matter expert.

A simulator script was developed for each scenario to assure that malfunctions were implemented identically for all control-room teams, and that the scenario ended as planned. Before starting the simulation, the control-room team was handed a standardised plant state briefing for each scenario, informing them about the initial operational state of the plant.

To gradually introduce the self-assessment for the operators, the control room teams firstly participated in two relatively brief scenarios planned to last about 20 min (scenario 1 and scenario 2). Thereafter, the teams participated in the remaining four scenarios planned to last for up to 45 min (scenario 3 to scenario 6).

Table 3
Overview of events of scenario 1 and scenario 6.

Scenario	Scenario event	Description of expected team behaviour
1	Loss of one busbar powered from the offsite system (loss of ordinary power of one safety train).	Team checks that emergency diesel generator starts, and that safety components start in sequence.
	Loss of all main feedwater pumps and malfunction of reactor cooling trains (loss of function needed for normal operation).	Team performs plant overview procedures for quick overview of plant status, interprets status and choses a strategy. Handles failures and prioritises to get all cooling trains operable.
6	One of the two condensate pumps stops, and the back-up pump starts.	Team detects the change of condensate pump and changes the combination of operating pumps.
	A leakage in an intermediate cooling system occurs (threatening normal operation).	Detects alarm, and orders field operator to fill up the system.
	Loss of offsite power and transition to house turbine power.	Detects loss of offsite power and detects power reduction; verifies house turbine operation.
	Air leakage to the condenser (threatening normal operation).	Team detects increasing pressure in the turbine condenser, mitigates the failure by starting back-up air ejector. Team is supposed to discuss the consequences and a strategy if turbine stop occurs.
	Loss of condenser; turbine and reactor stop. Loss of house turbine. Emergency diesel generators and safety components start.	Performs the plant overview procedures for a quick overview of plant status; interprets status and choses a strategy according to the plant status.
	Leakage in intermediate cooling system causes a shut-down pump malfunction.	Interprets that the pump stops owing to lost intermediate cooling system.
	Malfunctions in two of four emergency trains.	Team understands that two safety trains are not ready for operation; identifies and handles malfunctions.

3. Results

3.1. Simulation times and assessment procedure

The relatively complex simulation was completed without deviations from the planned scenario progressions. All malfunctions were implemented according to the simulator scripts for all crews and all scenarios. The crews' mean simulation times in minutes for scenario one to scenario six were as follows (the range is provided within the parentheses): 16:57 (12:33–24:48), 19:58 (12:34–26:24), 42:12 (37:16–48:16), 53:39 (45:53–62:26), 48:18 (43:02–56:54), and 32:42 (28:22–36:25). The variability in simulation time between crews reflects interaction between crews' performance and the dynamic development of the plant process, as well as variations in work style.

Regarding the self-assessment of performance, the operators quickly became familiar with navigating the replay tool and the performance assessment software. After a few requests for assistance during the assessment of the first scenario, operators frequently assessed the following scenarios without any requests for assistance.

3.2. Observation items and assessment items completed by the assessors

The assessors judged if an item was applicable or not and judged if the item could be assessed from the video and simulator data. For example, an item would be "not applicable" if the crew skipped a task, and an item could be "not assessable" due to cluttered communication. Consequently, the assessors completed a different number of items, based on their judgements.

Table 4 shows the ratio of *observation* items completed by the assessors as well as the range across teams. For the experts, the ratio is a mean across all teams, while the range provides the experts' minimum and maximum ratios of the six teams assessed. For example, expert 1 completed on average for all teams 0.96 of the monitoring items, while expert 1's lowest completion ratio for any team was 0.94 and the highest was 0.98. For the operator positions (supervisor, reactor operator, turbine operator) the ratio represents a mean of the six operators, and the range provides the lowest and highest proportion completed among the six teams within the given position. For example, for supervisors assessing monitoring, the mean ratio was 0.90 while the lowest completion among supervisors was .65 and the highest was 0.97. The "All Assessors" column of Table 4 provides the average ratio of items completed by teams' five assessors (both experts and the three operator

Table 4
Ratio of completed Observation items by the Assessors. Range across teams provided in parentheses.

Dimension	Items per team	Expert 1	Expert 2	Supervisors	Reactor Ops	Turbine Ops	All Assessors
Monitoring	63	.96 (.94,.98)	.96 (.92,1.0)	.90 (.65,.97)	.92 (.78,.98)	.82 (.73,.90)	.66 (.48,.86)
Strategy	5	.70 (.60,.80)	.90 (.80,1.0)	.67 (.40,.80)	.70 (.40,1.0)	.47 (.20,.80)	.20 (.00,.60)
Action	16	.81 (.69,.94)	.90 (.81,.94)	.83 (.75,.94)	.84 (.63,.94)	.81 (.69,.94)	.53 (.38,.81)
Verification	5	.93 (.80,1.0)	1.0 (1.0,1.0)	.97 (.80,1.0)	.87 (.60,1.0)	.57 (.40,.80)	.47 (.20,.80)
Total	89	.92 (.90,.94)	.95 (.92,.99)	.88 (.67,.96)	.89 (.75,.97)	.78 (.70,.88)	.60 (.46,.79)

positions). The range provides the lowest and highest ratio of jointly completed items among the six teams.

Table 4 shows that the experts completed the highest number of observation items in total. However, across dimensions, the number of items completed by the supervisors and reactor operators are quite similar to those of the experts. Turbine operators completed fewer items than the other assessors, except for action items. Overall, the turbine operators completed fewer items, probably because their competence profile differs from the other assessors. While the turbine operators possess general knowledge of the plant's reactor side, their competence on this part of the plant was limited, compared to that of the other assessors. The ranges provided in Table 4 show that there were individual differences among the operators and that, among the teams, there was at least one supervisor and one reactor operator who completed a similar number of items as the experts did.

The teams' assessors jointly judged an average ratio of 0.60 observation items as applicable and assessable, and the items completed by all assessors for each team were selected for analysing agreement among all assessors. This resulted in a data set consisting of 320 items (89 items * 6 teams * 0.60) assessed both by experts and by the respective operators within each of the six teams.

Table 5 shows the ratio of completed *assessment* items by the assessors as well as the ranges across teams. Overall, expert 2 completed the highest number of assessment items, the supervisors completed the second highest number, and expert 1 and the reactor operator completed a similar number of items, slightly below the number completed by the supervisors. In total, the turbine operators completed fewer assessment items than the other assessors, possibly because the competence profile of the turbine operators differs from the other control-room positions. The ranges provided show that there were individual differences among the operators, and that among the teams, there were examples of operators completing a similar number of items as the experts. Across performance dimensions, as with the observation items, the supervisors' and reactor operators' proportion of completed assessment items did not differ substantially from the two experts' proportion of completed items. The turbine operators consistently completed a lower proportion of items across all performance dimensions, except for strategy for which the completion ratio was like expert 1.

An average ratio of 0.61 assessment items was judged as applicable and assessable jointly by the teams' assessors, and these were selected for analysing agreement among all assessors. This resulted in a data set consisting of 894 items (245 items * 6 teams * 0.61) assessed by both experts and by the respective operators within each of the six teams.

3.3. Agreement on observations

The agreement among assessors for observation items was evaluated using Cohen's kappa (Cohen, 1960; Fleiss, 1971). There were no overarching observation items for interpretation, teamwork, goal, and work process. The observations items included the performance dimensions monitoring, strategy, action and verification. However, too few strategy and verification items were completed by all assessors to calculate a meaningful kappa coefficient. Therefore, the analysis includes separate agreement for the monitoring and actions dimensions only.

Comparison of the operators' assessment to experts' assessment interpretation of kappa coefficients was based on Landis and Koch (1977), suggesting values of 0.21–40 as "fair agreement", 0.41 to 0.60 to indicating "moderate agreement", values of 0.61 to .80 indicating "substantial agreement", and 0.81 to 1.0 indicating "perfect agreement". This classification is frequently applied in research (Hallgren, 2012), but

Table 5
Ratio of Assessment items completed by the Assessors. Range across teams provided in parentheses.

Dimension	Items per team	Expert 1	Expert 2	Super-visors	Reactor Ops	Turbine Ops	All Assessors
Monitoring	76	.94 (.91,.99)	.99 (.95,1.0)	.95 (.91,.99)	.94 (.80,1.0)	.82 (.74,.95)	.72 (.59,.91)
Interpretation	25	.85 (.68,.92)	.98 (.92, 1.0)	.93 (.76,1.0)	.92 (.80,1.0)	.61 (.44,.80)	.50 (.32,.68)
Strategy	9	.56 (.44,.67)	.81 (.78,.89)	.81 (.67,1.0)	.76 (.33,1.0)	.57 (.33,1.0)	.33 (.22,.44)
Action	20	.83 (.75,.90)	.98 (.90,1.0)	.86 (.75,.95)	.83 (.55,1.0)	.66 (.45,.90)	.44 (.25,.65)
Verification	31	.94 (.90,.97)	.98 (.90,1.0)	.92 (.81,1.0)	.92 (.71,1.0)	.75 (.55,.90)	.63 (.48,.74)
Teamwork	42	.87 (.76,.95)	.99 (.95,1.0)	.95 (.86,1.0)	.92 (.79,1.0)	.81 (.67,.93)	.69 (.62,.79)
Goal	18	.89 (.83,.94)	.96 (.83,1.0)	.97 (.94,1.0)	.82 (.50,1.0)	.67 (.33,.94)	.48 (.22,.89)
Work Proc.	24	.92 (.92,.92)	.92 (.92,.92)	.92 (.92,.92)	.76 (.21,1.0)	.69 (.50,.92)	.55 (.17,.92)
Total	245	.89 (.87,.91)	.97 (.96,.99)	.93 (.87,.96)	.89 (.67,.99)	.74 (.63,.90)	.61 (.46,.80)

it can be noted that more conservative guidance is available (Krippendorff, 1980).

Table 6 shows the kappa coefficient and the 95% confidence interval for each pair of assessors. To ease the overview of the relatively large number of pairwise agreements, the "perfect" agreement has been coded with a light green background and a solid line below the kappa coefficient; "substantial" agreement has been coded with light yellow background with a dotted line. "Fair" and "moderate" agreement have no coding.

Overall, the pairs of assessors agreed substantially, except for three of the four pairs of assessors including the turbine operators. The pair of experts agreed perfectly on actions while the expert-operator agreement was substantial. For monitoring, the supervisors' agreement with both experts and reactor operator-expert 2 agreement were classified as substantial and identical to the expert agreement. Within the control-room team, supervisors and reactor operators agreed substantially. The agreement between supervisors and turbine operators were low. Interestingly, all assessor pairs of agreement for action items were higher than for monitoring items, except for the supervisor-turbine operators' agreement, which was identical for monitoring and actions.

3.4. Agreement on assessment

The agreement among assessors for assessment items was evaluated using the intraclass correlation (ICC) (Shrout and Fleiss, 1979). The ICC model applied was one-way random effects, absolute agreement, and two assessors per item. For the pair of experts, a two-way model would also be appropriate since the experts rated the items for all crews. However, for comparability of ICC results, a one-way model was run for all pairs of assessors. Table 7 shows the ICC and the 95% confidence interval for each pair of assessors.

To compare the operators' assessment with the experts' assessment, interpretation of ICC was based on the classification provided by Cicchetti (1994), classifying agreement as "poor" for values less than 0.40, "fair" for values between 0.40 and 0.59, "good" for values between 0.60 and 0.74, and "excellent" for values above 0.75. However, Koo and Li (2016) provides a more conservative interpretation of ICC, which might be considered for other applications.

Table 7 presents the ICC and the 95% confidence interval for each pair of assessors. To ease the overview of the relatively large number of pairwise agreements, "excellent" agreement is coded with a light green background and a solid line below the ICC point estimate; "good" agreement is coded with light yellow background with a dotted line. "Poor" and "fair" agreement have no coding.

Overall, the experts' agreement can be interpreted as one level above the operator-expert agreement. With regard to overall assessment, the experts' agreement was classified as "excellent". The operator-expert agreement was classified as "good", as was the agreement within the control-room team. However, for strategy the operator-expert agreement was similar to the experts' agreement and in the case of the agreement on the dimensions action and teamwork, several of the pairs of operator-expert agreement were similar to the experts' agreement.

With regard to the agreement of individual performance dimensions, it is noteworthy that experts showed "excellent" agreement on monitoring, while operator-experts showed less agreement. For monitoring, most operator-expert agreement was "good". Experts' agreement on

Table 6
Agreement of Cohen's Kappa for each pair of Assessors. 95% confidence intervals provided below each Kappa coefficient.

Assessor pair		All items	Monitoring	Action
Expert 1 -	Expert 2	.73	0.66	0.82
		(.63,.84)	(.51,.81)	(.65,.99)
Supervisor -	Expert 1	.66	.62	.73
		(.54,.78)	(.45,.78)	(.52,.93)
	Expert 2	.69	.70	.72
		(.58,.80)	(.56,.84)	(.50,.93)
Reactor Op.-	Expert 1	.62	.54	0.77
		(.50,.75)	(.37,.71)	(.58,.96)
	Expert 2	.66	.63	.76
		(.54,.78)	(.48,.78)	(.56,.96)
Turbine Op.-	Expert 1	.61	.50	.74
		(.48,.74)	(.31,.69)	(.54,.94)
	Expert 2	.60	.49	.74
		(.47,.72)	(.32,.66)	(.53,.94)
Supervisor -	Reactor Op	.71	.72	.75
		(.60,.83)	(.57,.86)	(.54,.96)
	Turbine Op	.55	.55	.55
		(.40,.69)	(.36,.74)	(.30,.80)
Reactor Op.-	Turbine Op	.60	.52	.68
		(.47,.74)	(.34,.71)	(.46,.90)

Table 7
ICC for Assessment items for each pair of Assessors. 95% confidence intervals are provided in parentheses below each Kappa coefficient.

Assessor pair		Tot.	Mon.	Int.	Strat.	Action	Verif.	Teamw.	Goals	Workp.
Expert 1	Exp.2	.75	.77	.62	.92	.84	.61	.80	.47	.12
		(.72,.78)	(.71,.81)	(.39,.76)	(.80,.97)	(.73,.91)	(.44,.73)	(.73,.85)	(.09,.70)	(.--,.44)
Superv.	Exp.1	.70	.65	.76	.94	.75	.42	.77	.29	.54
		(.65,.73)	(.57,.72)	(.62,.85)	(.83,.98)	(.57,.86)	(.16,.60)	(.69,.83)	(.--,.59)	(.28,.71)
	Exp.2	.65	.61	.67	.88	.82	.22	.75	--.30	.28
		(.60,.69)	(.52,.69)	(.48,.79)	(.69,.96)	(.69,.90)	(.--,.45)	(.67,.82)	(.--,.25)	(.--,.5)
Reac.Op.	Exp.1	.64	.62	.72	.78	.83	.35	.72	--.12	.35
		(.59,.68)	(.53,.69	(.55,.82)	(.42,.92)	(.71,.90)	(.07,.55)	(.62,.79)	(.--,.36)	(.--,.59)
	Exp.2	.66	.63	.60	.87	.81	.39	.75	.52	.35
		(.62,.71)	(.54,.70)	(.37,.75)	(.65,.95)	(.68,.89)	(.13,.58)	(.66,.81)	(.17,.72)	(.--,.58)
Turb.Op.	Exp.1	.60	0.49	.60	.85	.71	.38	.71	.52	.29
		(.54,.64)	(.37,.59)	(.37,.75)	(.62,.94)	(.50,.83)	(.10,.57)	(.61,.79)	(.17,.73)	(.--,.55)
	Exp. 2	.63	.52	.55	.87	.78	.36	.77	.67	.32
		(.58,.68)	(.41,.62)	(.29,.72)	(.67,.95)	(.63,.88)	(.08,.56)	(.69,.83)	(.43,.81)	(.--,.56)
Superv.	R.Op.	.66	.60	.75	.91	.76	.55	.71	.15	.20
		(.61,.70)	(.51,.68)	(.60,.84)	(.77,.97)	(.59,.86)	(.35,.68)	(.61,.79)	(.--,.51)	(.--,.48)
	T.Op.	.63	.49	.68	.90	.66	.57	.75	.13	.18
		(.57,.67)	(.37,.59)	(.50,.80)	(.73,.96)	(.42,.80)	(.38,.70)	(.67,.82)	(.--,.50)	(.--,.47)
Reac.Op.	T.Op.	.67	.54	.65	.92	.72	.72	.69	.62	.54
		(.62,.71)	(.43,.63)	(.44,.78)	(.78,.97)	(.51,.84)	(.60,.81)	(.59,.77)	(.33,.78)	(.29,.71)

interpretation was "good", while operator-expert agreement was either similar or better, except for the turbine operator-expert 2 agreement. Experts' agreement regarding verification was "good" only, while operator-expert agreement was "poor" or "fair". Agreement on assessing goals and work process was low among all assessor pairs, except for the turbine operator-expert 2 regarding goals.

3.5. Illustrating agreement by three assessor configurations

A commonly applied evaluation of measurement reliability is the agreement among assessors. To illustrate agreement among experts and operators in a reliability context, the ICC for three different assessor configurations was calculated. The three configurations consisted of the two experts, the three operator positions from each of the six teams, and both the experts and the operators, respectively. For a given team, and thereby for each item included in the analysis, two experts (n = 2) and the team's three operators (n = 3) assessed performance. However, in total, eighteen operators (n = 18) made up the three operator positions for the six teams. Fig. 4 shows the ICC and the 95% confidence interval for the three configurations.

The assessor configuration of the two experts and the configuration of the three operator positions resulted in similar level of agreement, ICC of 0.75 and 0.74 respectively, even though the operator configuration consisted of three assessors for each item while there were two experts. As shown in Table 7 above, the agreement among each pair of three operator positions was lower than the agreement between the two experts. Adding operators to the experts as assessors significantly increased reliability from 0.75 to 0.83, $F(898,3596) = 1.48$, $p < .001$. Adding operators as additional assessors influenced ICC in the direction as expected when adding assessors with ICC <1 to the original assessors (Kahan et al., 2017).

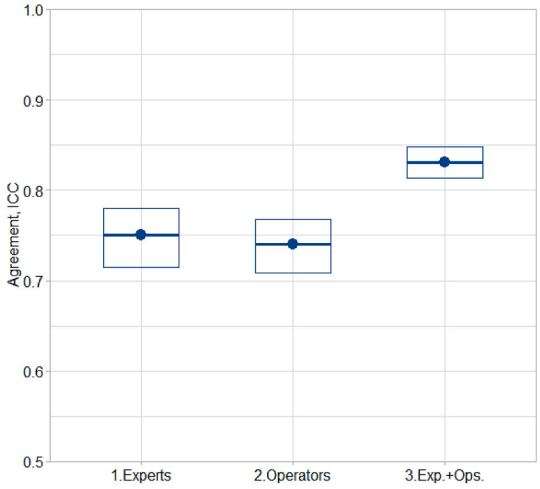


Fig. 4. Reliability of different assessor configurations.

3.6. Assessors' judgement of performance level

In addition to exploring the agreement and measurement reliability, the study investigated the performance level resulting from the expert's assessment of the operator's self-assessment. Fig. 5 shows the calculated mean performance level for each assessor type across the SCORE performance dimensions.

Owing to the unbalanced number of cases across the performance dimension and the different number of assessors within each assessor category, a linear mixed model was used to investigate the effect of assessor category and performance dimension on the rated performance level (West et al., 2015). The results of the linear mixed model showed that there was no significant main effect of assessor category, $F(4, 15) = 0.65$, $p = .63$, and no significant interaction effect between assessor category and performance dimension, $F(28, 4441) = 0.96$, $p = .53$. There was a significant effect of performance dimension, $F(7, 4441) = 22.29$, $p < .001$, $\eta = 0.13$. Post-hoc comparisons using marginal means, applying Tukey HSD adjustment, showed that monitoring was rated significantly different from the other dimensions, apart from strategy,

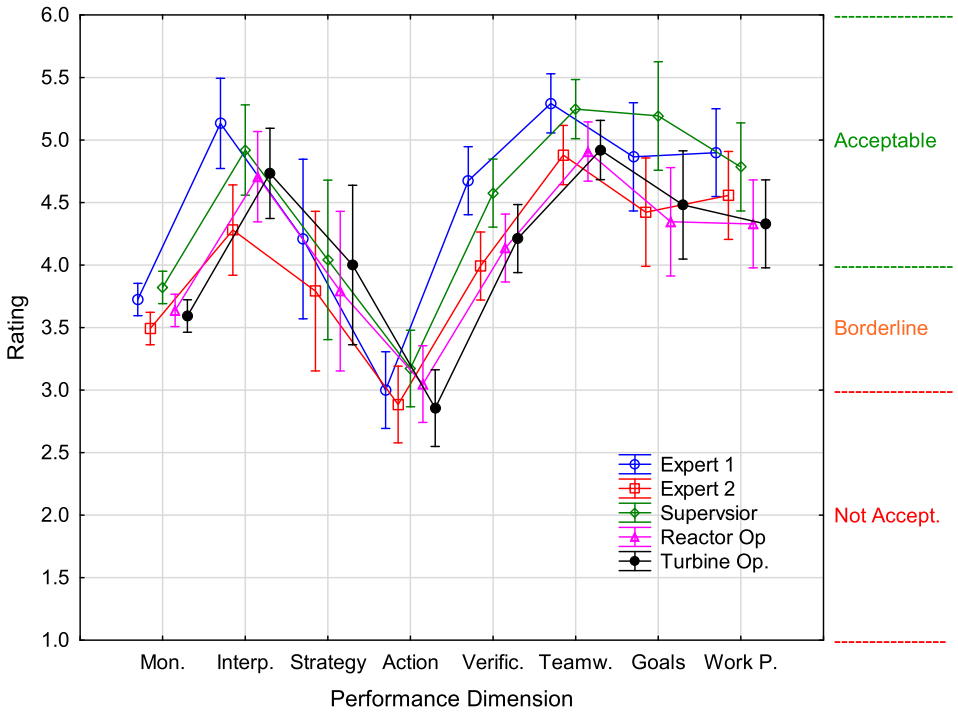


Fig. 5. Assessors' performance level. Mean for each assessor type. Vertical bars denote 0.95 confidence intervals. The labelling of the right y-axis is abstracted from the rating scale anchors.

and action was rated significantly lower than all other dimensions (all adjusted p-values $< .05$).

3.7. Expert assessment versus operator self-assessment for the identification of non-acceptable performance items

Beyond the overall agreement between experts and operators, items with a low rating are of interest for identifying improvements to a human-machine system. To compare observer assessment versus self-assessment for this type of identification, all items assessed "not acceptable" (rated as 1 or 2) by any of the assessors were selected. A total of 177 items were scored as "not acceptable" by any one of the assessors. Table 8 shows the proportion of items identified by experts only, both experts and operators, or operators only for each performance dimension, and in total.

Overall, the experts alone identified 0.19 of the items scored as "not acceptable" by an assessor, while the operators alone identified 0.49 of these. The operators consistently identified a higher ratio of unacceptable items across all performance dimensions, except for action items where assessors jointly identified 0.67 and operators alone identified a ratio of 0.19. The operators' identification ratio was especially high for goals and for work process: 0.78 and 0.93 respectively.

Table 8
Proportion of low-score items identified by Assessor type.

		Proportion of low score items identified by Assessor type		
Performance Dimension	n	Experts only	Experts & Operators	Operators Only
Monitoring	47	.21	.38	.40
Interpretation	20	.25	.30	.45
Strategy	5		.60	.40
Action	21	.14	.67	.19
Verification	29	.24	.14	.62
Teamwork	31	.19	.39	.42
Goals (Overall goal)	9	.22		.78
Work Process	15		.07	.93
Sum	177	.19	.33	.49

3.8. Interview

A majority of the operators found the self-assessment was enjoyable and useful, and none expressed any discomfort or negative feelings regarding the self-assessment. All Operators reported that the SCORE items and the scenarios were relevant for assessing control-room work, and the majority of the operators felt that this type of self-assessment could be relevant for their simulator-based training. For example, five of the six teams noted that the study almost represented "an additional re-training". Two teams mentioned that self-evaluation could help reflect on one's own behaviour. Further, self-evaluation could be a more neutral evaluation than external evaluation, which tended to focus more on the negative observations.

Four of the six turbine operators stated that it was difficult to rate the items pertaining mostly to the supervisor or reactor operator position because their competence was concentrated mainly on the plant's turbine side. It was especially challenging to evaluate items pertaining to the supervisor. None of the supervisors or reactor operators experienced difficulties in assessing items involving the turbine position.

4. Discussion

This study provided participating operators and experts with the same scenario replay tool, the same event specific assessment protocol, and similar assessment procedures to investigate self-assessment as compared to expert assessment. Under these conditions, the study found overall agreement between operators and experts, close to the agreement found among experts. For assessment items, the agreement between two experts, measured by the ICC, was 0.75 and classified as "excellent". The agreement between operators and experts ranged from 0.60 to 0.70 and was classified as "good". The observed agreement was comparable to the average expert observer agreement reported in a literature review of the medical domain at about 0.70 (Mete and Brannick, 2017) and in aviation at about 0.75 (Holt et al., 2002). In addition,

the operators' self-assessment provided information beyond the expert assessment in identifying "non-acceptable" performance items. Of all performance items assessed as "non-acceptable" by either experts or participants, 49% of the items were identified by self-assessment only. In the interview, the operators reported that they found the self-assessment enjoyable. They confirmed the relevance of the SCORE items for assessing control room work and considered this type of self-assessment relevant for training. Furthermore, the majority of turbine operators expressed that, due to their area of competence, it was challenging to rate the items pertaining mostly to the supervisor or reactor operator positions.

4.1. Assessors' completion of items

The assessors' ratios of completed items suggest that operators and experts employ similar judgements regarding the applicability of scenario-specific items to given scenario progressions, given that the operators have similar competence profiles as the experts. The supervisors and reactor operators completed similar numbers of observation items and assessment items as the experts did (both in total and for most individual performance dimensions), while the turbine operators completed a lower number of items. Results from the interviews confirm that the competence profile for turbine operators is a probable explanation for the lower number of completed items. Turbine operators' main technical competence is concentrated on the turbine systems, while the competence of both supervisors and reactor operators covers both the reactor systems and the turbine systems.

4.2. Accuracy of operators' self-assessment

The overall assessment results support the utility and validity of professional operators' self-assessment of complex dynamic work. With regard to *observation* items (which were a secondary supportive element for the assessment), the operator-expert agreement regarding an item performed or not performed by the team was similar to the respective agreement among the experts. Expert agreement, measured by Cohen's kappa, was 0.73, while the pairwise operator-expert agreement ranged from 0.60 to 0.69. Both expert agreement and operator-expert agreement were classified as "substantial", except for the turbine operator-expert 2 agreement, which was classified as "moderate".

Overall, the agreement between operators' and experts' ratings of *assessment* items was moderately lower than the agreement found among experts. Expert agreement, measured by the ICC, was 0.75, while the pairwise operator-expert agreement ranged from 0.60 to 0.70. The agreement was classified as "excellent" and "good", respectively. However, the results include examples of the performance dimension with operator-expert agreement similar or very close to the experts' agreement. The high agreement on strategy, action, and teamwork were such examples, as was the poor agreement on goals and work process. These results suggest potential for achieving a similar level of agreement with operators as with experts.

The overall ICCs suggest that both groups of assessors would provide overall results of similar reliability. The overall ICCs for experts ($n = 2$) and operator positions ($n = 3$) were 0.75 and 0.74, respectively. Decisions regarding assessors can then be based on practical constraints rather than concerns about attaining different levels of reliability. Additionally, using all assessors significantly increased the reliability in terms of the ICC to 0.85. This suggests that adding self-assessment to observer assessment could increase measurement reliability towards a targeted level. In this respect, the results support the general expectation that adding additional assessors increases reliability (Kahan et al., 2017; Mete and Brannick, 2017).

4.3. Operators' versus experts' assessment of overall performance level

Beyond measurement reliability, the study suggests that using either experts or operators to assess overall performance would yield similar conclusions. As could be partly expected from the relatively high agreement observed between experts and operators, the study did not identify a significant difference between the assessors' overall performance assessment. The mean performance rating of all assessors was in a similar region of the scale (acceptable or borderline as presented in Fig. 5) for all performance dimensions, even though there was some variation in the ratings between the two experts and between the control room positions. Contrary to studies reporting low accuracy for determining acceptable performance or pass versus fail (O'Connor et al., 2002; Gontar and Hoerman, 2015), this study suggests potential for utilizing operator self-evaluation to determine acceptable performance levels. However, in addition to substantial agreement, the results also indicate an operator perspective on performance, identifying performance issues not captured by expert assessment. Investigating the low score items demonstrated that 49% of the items assessed as non-acceptable were identified by the self-assessment, while only 19% were identified by the expert assessment. This additional perspective represents critical evaluation of one's own performance rather than overconfidence (Davis et al., 2006; Ehrlinger et al., 2008).

4.4. Specificity and observability of performance dimensions

A hypothetical explanation for differing levels of agreement across performance dimensions is based on the item's task specificity (Gaba et al., 1998; Brannick et al., 2002; Nixon et al., 2015) and observability, such as cognitive versus overt behavioural or system indicators (Flin and Martin, 2001; Gontar and Hoermann, 2015). "High" agreement dimensions (strategy, action, and teamwork) included specific items for which overt behaviour or human-system indicators could support the assessment. "Moderate" agreement dimensions (monitoring and interpretation) included a high number of specific cognitive items for which overt behaviour or human-system indicators could indirectly support the assessment. One type of "poor" agreement dimension (verification) included specific cognitive items with limited overt human-system indicators, and a second type of "poor" agreement dimensions (goals and work process) included items that were general in nature. For example, strategy items were usually discussed and announced within the team and the scenario replay provided a basis for the assessment of these items. Action items included controlling plant systems, and the replay tool provided an overview of the teams' implementation and timing of main actions. In this study, teamwork was mainly the communication of key technical work to coordinate and update the team. The operators' recorded speech provided a solid basis for assessing these items.

In contrast, the content of monitoring and interpretation was also specific but, to a higher degree, purely cognitive with fewer direct indicators. For monitoring and interpretation, an assessor could observe if the plant condition for the behaviour was present (e.g., that the event resulted in key alarms that should be detected and interpreted). However, the assessor needed to perform a judgement based on observable team behaviours (whether an item was performed or not) and to assess the quality of the performance (Flin and Martin, 2001; O'Connor et al., 2002). For verification, the cognitive behaviours addressed were usually initiated by the operator rather than a human-machine event; thereby, limited human-system indicators were present to support the assessment. Regarding goals and work process, the content of these items was relatively general. For example, goals concerned a global evaluation of the team's overview of an event and handling of the event's malfunctions. Work process concerned global evaluation of the management and utilisation of human-machine interfaces during the scenario. The low accuracy of general items, as well as the relatively high proportion of non-acceptable performance operators identified for these items, might possibly be due to assessors attending to different aspects of behaviour (Weber et al., 2013) and integrating indicators differently for a global assessment (Gaba et al., 1998).

Importantly, the hypothetical explanation above applies to both

expert and operator assessment. The pattern of agreement across dimensions was relatively similar for expert agreement and operator-expert agreement. For example, both expert agreement and operator-expert agreement were generally high on the "high" agreement dimensions and low on the "low" dimensions of goals and work process. Presumably, task specificity and observability of items influenced experts' and operators' assessment in a relatively similar manner, which is an argument supporting operator assessment as a potentially valid alternative and complement to expert observer assessment. The substantial agreement on cognitive items (monitoring and interpretation) corresponds with results from aviation (Gontar et al., 2014). However, for cognitive behaviours related to limited observable indicators (verification) this study suggests relatively less agreement between operators' self-assessment and experts' assessment.

4.5. The assessment method applied in the study

This study contributes to the literature by using the same task-specific measure, assessment tool, and assessment procedure for the comparison of operators' and experts' performance assessment. Frequently, this is not the case. For example, expert observers monitor work and perform assessment concurrently, while participants apply self-assessment post scenario, based on their memory of the event (e.g., Andrew et al., 2012; Weller et al., 2013; Andersson et al., 2017). One could hypothesize that the study's replay tool influenced agreement, at least as reported for video (Martin et al., 1998; Ward et al., 2003; Nayar et al., 2020), since the tool provided full log and dynamic replay of system behaviours and interface activities in addition to video and audio recordings of the team's work. However, regarding the poor agreement observed for the general items (goals and work process), replay of the scenario might not in itself be sufficient to achieve high agreement (Gaba et al., 1998). In complex scenarios, several tasks are often interwoven. For example, an operator might be monitoring the dynamic process, developing a hypothesis about the situation, and communicating with teammates in between these two actions. The task-specific measure may have supported the operators to attend to the same key aspects of the performance as the experts attended to, thereby holding a basis for their assessment that is comparable to those of the experts. In the interview, the operators' reporting of the SCORE items' relevance for assessing control room work makes it plausible that this type of task-specific measure substantially guided their assessment.

The procedure of informing the operators about the failures implemented in the scenario (after performing the scenario but before the self-assessment) provided the operators with the actual system events to evaluate their performance against. In cases where a failure was not detected or was misinterpreted by the team (e.g., due to masked plant information), knowing what failure occurred may have had substantial influence on an operator's evaluation of their own performance.

One can hypothesize whether agreement could be increased by further improving the assessment method. The main best practice not sufficiently implemented in this study was training the assessors (Rosen et al., 2008; Mete and Brannick, 2017). Training the assessors consisted of an oral explanation of the SCORE method and facilitating a relatively simple example while demonstrating the scenario replay and the assessment software. Running a realistic test self-assessment, followed by feedback and a group discussion on the performed assessment could provide an opportunity for clarification and an increased understanding of the assessment method. Providing a common frame of reference (Bernardin and Buckley, 1981) could also increase the shared understanding of the intended performance assessment.

The turbine operators – completing a relatively low proportion of items and performing less accurate assessment than the supervisors and reactor operators – demonstrated the importance of assessor competence. The results involving the turbine operators correspond to Yule et al.'s (2008) results regarding accuracy of surgeons' performance assessment which depended on a match between surgeon's speciality and the tasks to be evaluated. For interdisciplinary teams of specialists, assessor competence related to the team's set of tasks is a relevant issue for the assessment procedure. In this study, an alternative procedure could have been turbine operators assessing their team's tasks as related to their specific competence only.

4.6. Limitations and future research

The study included a total of 18 male operators from six control room teams, and two expert assessors. This is a relatively small sample and relates to the practical challenge of recruiting professional control room operators and experts for simulator studies of human-machine industry systems. Furthermore, control room work in nuclear process control has certain characteristics that may relate to the operator-expert agreement observed. Nuclear power plants are often one-of-a-kind human-machine designs with specific interfaces and operating procedures and a distinct concept of operation. Several years of plant-specific education and simulator training are required to qualify as an operator. Prospective subject matter experts on control room work are frequently recruited from the population of these specialised operators. As such experts and operators have much core control room competence in common, both types of assessors classify as substantially competent, a characteristic related to relatively high self-assessment accuracy (Kruger and Dunning, 1999; Ehrlinger et al., 2008). Future research should investigate replicability of the results in process control and investigate to what extent these findings extend to other settings.

The study did not include a control group applying a simple generic performance protocol (with or without scenario replay) or a control group applying the assessment protocol from memory of the scenario (without scenario replay). Future research could investigate operators' task-specific assessment, both with and without scenario replay, as compared to expert assessment utilizing the scenario replay tool.

4.7. General implications

The validity of self-assessment is a critical assumption for its application in evaluating complex human-machine systems, as well as for its application in professional competence development. The study's results suggest potential for achieving relatively accurate self-assessment in complex dynamic work settings. The study suggests that using task-specific assessment items developed by subject matter experts or similar professionals and using some form of scenario replay are important for self-assessment accuracy. Under such conditions, self-assessment of accuracy similar to that of expert assessment can be achieved, and self-assessment can capture performance aspects not easily observable (e.g., the assessment of cognitive behaviour).

One can expect that new technology provides capabilities for efficient implementation of self-assessment procedures and scenario replay similar to those applied in this study. New approaches can effectively combine video, audio, interface, and system behaviour for event-based assessment, and this can be integrated with assessment protocols and assessment procedures. The enjoyment of self-assessment reported by the operators suggest a motivational aspect important for self-regulated learning and self-regulated improvements in working conditions. In these conditions, accurate self-assessment can augment human-system evaluation and competence development (Nayar et al., 2020) and can thus reduce demand for expert resources (Arora et al., 2011).

5. Conclusion

By providing operators with an assessment method in line with recommended practice, this study demonstrates a possibility for achieving relatively high accuracy of operators' self-assessment for complex dynamic work. In addition, self-assessment might yield insights not easily captured by expert observers. Future research could investigate to what extent scenario replay, task specificity, and observability of

measurement items impact self-assessment accuracy and what type of additional insights can be gained from self-assessment as compared to expert assessment.

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 author would like to thank an anonymous reviewer whose comments greatly helped to improve and clarify this manuscript.

This work was funded by the OECD Halden Reactor Project, Norway.

References

Andersson, D., Rankin, A., Diptee, D., 2017. Approaches to team performance assessment: a comparison of self-assessment reports and behavioral observer scales. Cognit. Technol. Work 19, 517–528.

Andrew, B., Plachta, S., Salud, L., Pugh, C.M., 2012. Development and evaluation of a decision-based simulation for assessment of team skills. Surgery 152 (2), 152–157. https://doi.org/10.1016/j.surg.2012.02.018.

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

Arora, S., Miskovic, D., Hull, L., et al., 2011. Self vs expert assessment of technical and non-technical skills in high fidelity simulation. Am. J. Surg. 202 (4), 500–506, 2011.

Bernardin, H.J., Buckley, M.R., 1981. Strategies in rater training. Acad. Manag. Rev. 6, 205–212.

Boud, D., 1995. Assessment and learning: contradictory or complementary assessment for learning in higher education. In: Knight, P. (Ed.), Assessment for Learning in Higher Education. Kogan Page, London, pp. 35–48.

Boud, D., Lawson, R., Thompson, D., 2015. The calibration of student judgement through self-assessment: disruptive effects of assessment patterns. High Educ. Res. Dev. 34 (1), 45–59. https://doi.org/10.1080/07294360.2014.934328.

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., Berntsson, O., 2016. Assessment of situation understanding, mission, control and teamwork in the control room: the development and initial testing of the SCORE measure. HWR-1125. Halden, Norway. OECD Halden Reactor Project.

Braarud, P.Ø., Svengren, H., 2020. *Evaluation of first check Procedure and technical specification overview – Comparing computerized and paper-based solutions* (halden work report 1299). OECD Halden Reactor Project, Halden, Norway.

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., Svengren, H., Hunton, P., Joe, J., Hanes, L., 2019. A graded approach to the human factors validation of turbine control system digital upgrade and control room modernization. Proceedings of the Eleventh American Nuclear Society International Topical on Nuclear Plant Instrumentation, Controls and Human-Machine Interface Technologies (NPIC & HMIT 2019). American Nuclear Society, Orlando, FL.

Brannick, M.T., Prince, C., Salas, E., 2002. The reliability of instructor evaluations of crew performance: good news and not so good news. Int. J. Aviat. Psychol. 12 (3), 241–261. https://doi.org/10.1207/S15327108IJAP1203_4.

Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess. 6 (4), 284–290.

Cohen J. (1060). A coefficient of agreement for nominal scales. Educ. Psychol. Meas.. 20: 37–46.

Colthart, I., Bagnall, G., Evans, A., Allbutt, H., Haig, A., Illing, J., et al., 2008. The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice: BEME Guide no. 10. Med. Teach. 30 (2), 124–145.

Davis, D.A., Mazmanian, P.E., Fordis, M., Van Harrison, R., Thorpe, K.E., Perrier, L., 2006. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. JAMA, J. Am. Med. Assoc. 296 (9), 1094–1102.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., Kruger, J., 2008. Why the unskilled are unaware: further explorations of (absent) self-insight among the incompetent. Organ. Behav. Hum. Decis. Process. 105 (1), 98–121. https://doi.org/10.1016/j.obhdp.2007.05.002.

Eitrheim, M.H.R., Svengren, H., Fernandes, A., 2018. Computer-based human-machine interfaces for emergency operation. Nucl. Technol. https://doi.org/10.1080/00295450.2018.1426962.

Eva, K.W., Regehr, G., 2005. Self-assessment in the health professions: a reformulation and research agenda. Acad. Med. 80 (10 Suppl. l), 46–54.

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychol. Bull. 76 (5), 378–382.

Flin, R., Martin, L., 2001. Behavioural markers for crew resource management: a survey of current practice. Int. J. Aviat. Psychol. 11, 95–118.

Gaba, D.M., Howard, S.K., Flanagan, B., Smith, B.E., Fish, K.J., Botney, R., 1998. Assessment of clinical performance during simulated crises using both technical and behavioral ratings. Anesthesiology 89 (1), 8–18.

Ganni, S., Chmarra, M.K., Goossens, R.H.M., Jakimowicz, J.J., 2017. Self-assessment in laparoscopic surgical skills training: is it reliable? Surg. Endosc. 31 (6), 2451–2456.

Gontar, P., Hoermann, H.J., 2015. Interrater reliability at the top end: measures of pilots' nontechnical performance. Int. J. Aviat. Psychol. 25 (3–4), 171–190. https://doi.org/10.1080/10508414.2015.1162636.

Gontar, P., Hoermann, H.J., Deischl, J., Haslbeck, A., 2014. How pilots assess their non-technical performance - A flight simulator study (Krakow). In: Stanton, N.A., Landry, S.J., Di Bucchianico, G., Vallicelli, A. (Eds.), Advances in Human Aspects of Transportation. Part I, pp. 119–128.

Gordon, M.J., 1991. A review of the validity and accuracy of self-assessments in health professions training. Acad. Med.: journal of the Association of American Medical Colleges 66 (12), 762–769. https://doi.org/10.1097/00001888-199112000-00012.

Gordon, M.J., 1992. Self-assessment programs and their implications for health professions training. Acad. Med. 67 (10), 672–679.

Gunnarsson, T., Braarud, P.Ø., Fernandes, A., 2014. Performance Based Control Room Evaluation (ISV) for Oskarshamn 1 Periodic Safety Review. Paper presented at Enlarged Halden Programme Group Meeting, EHPG 2014, Røros, Norway.

Hall, S., Brannick, M.T., 2009. Performance assessment in simulation. In: Vinncenzi, D. A., Wise, J.A., Mouloua, M., Hancock, P.A. (Eds.), Human Factors in Simulation and Training. CRC Press, pp. 149–168.

Hallgren, K.A., 2012. Computing inter-rater reliability for observational data: an overview and tutorial. Tutorials in quantitative methods for psychology 8 (1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023.

Holt, R., Hansberger, J., Boehm-Davis, D., 2002. Improving rater calibration in aviation: a case study. Int. J. Aviat. Psychol. 12 (3), 305–330. https://doi.org/10.1207/S15327108IJAP1203_7.

IFE, 2020. HAMMLAB. Institute for Energy Technology. https://ife.no/en/laboratory/hammlab/.

Kahan, B.C., Feagan, B., Jairath, V., 2017. A comparison of approaches for adjudicating outcomes in clinical trials. Trials 18 (1), 266. https://doi.org/10.1186/s13063-017-1995-3.

Kim, J.H., 2018. The effect of metacognitive monitoring feedback on performance in a computer-based training simulation. Appl. Ergon. 67, 193–202. https://doi.org/10.1016/j.apergo.2017.10.006.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of chiropractic medicine 15 (2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Krippendorff, K., 1980. Content Analysis: an Introduction to its Methodology. Sage Publications, Beverly Hills, CA.

Kruger, J., Dunning, D., 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J. Pers. Soc. Psychol. 77 (6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174.

Landy, F.J., Farr, J.L., 1983. The Measurement of Work Performance: Methods, Theory, and Applications. Academic Press, New York.

Laumann, K., Braarud, P.Ø., Svengren, H., Bye, A., 2006. Study on how time pressure and information load affects operators performance in accident scenarios. New Orleans, USA Proceedings of PSAM 8. https://doi.org/10.1115/1.802442.paper81. International Conference on Probabilistic Safety Assessment and Management.

Marienau, C., 1999. Self-assessment at work: outcomes of adult learners' reflections on practice. Adult Educ. Q. 49 (3), 135–146.

Marriage, B., Kinnear, J., 2016. Assessing team performance - markers and methods. Trends in Anaesthesia and Critical Care 7–8. https://doi.org/10.1016/j.tacc.2016.05.002.

Martin, D., Regehr, G., Hodges, B., McNaughton, N., 1998. Using videotaped benchmarks to improve the self-assessment ability of family practice residents. Acad. Med. 73, 1201–1206.

Mete, I., Brannick, M.T.(, 2017. Estimating the reliability of nontechnical skills in medical teams. J. Surg. Educ. 74 (4), 596–611.

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

Nayar, S.K., Musto, L., Baruah, G., Fernandes, R., Bharathan, R., 2020. Self-assessment of surgical skills: a systematic review. J. Surg. Educ. 77 (2), 348–361. https://doi.org/10.1016/j.jsurg.2019.09.016.

Nixon, J., Leggatt, A., Campbell, J., 2015. The development and assessment of behavioural markers to support counter-IED training. Appl. Ergon. 48, 130–137.

O'Connor, P., Hörmann, H.J., Flin, R., Lodge, M., Goeters, K.M., The JARTEL Group, 2002. Developing a method for evaluating crew resource management skills: a European perspective. Int. J. Aviat. Psychol. 12 (3), 263–285. https://doi.org/10.1207/S15327108IJAP1203_5.

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.

Rasmussen, J., 1986. Information Processing and Human-Machine Interaction: an Approach to Cognitive Engineering. North-Holland.

Rosen, M., Salas, E., Wilson, K.A., King, H.B., Salisbury, M., Augenstein, J.S., Robinson, D.W., Birnbach, D.J., 2008. Measuring team performance in simulation-based training: adopting best practices for healthcare. Simulat. Healthc. J. Soc. Med. Simulat. 3 (1), 33–41. https://doi.org/10.1097/SIH.0b013e3181626276.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86 (2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420.

Simonsen, E., Osvalder, A.L., 2018. Categories of measures to guide choice of human factors methods for nuclear power plant control room evaluation. Saf. Sci. 102, 101–109.

Sinclair, M.A., 1995. Subjective assessment. In: Wilson, J.R., Corlett, E.N. (Eds.), Evaluation of Human Work: A Practical Ergonomics Methodology, second ed. Taylor & Francis, London, pp. 69–100.

Steinemann, S., Berg, B., Ditullio, A., Skinner, A., Terada, K., Anzelon, K., Ho, H.C., 2012. Assessing teamwork in the trauma bay: introduction of a modified "NOTECHS" scale for trauma. Am. J. Surg. 203 (1), 69–75.

Van Loon, M.H., 2018. Self-assessment and self-reflection to measure and improve self-regulated learning in the workplace. In: McGrath, S., Mulder, M., Papier, J., Suart, R. (Eds.), Handbook of Vocational Education and Training. Springer, Cham, pp. 1–20.

Vicente, K.J., 1999. Cognitive Work Analysis: toward Safe, Productive, and Healthy Computer-Based Work. Lawrence Erlbaum Associates Publishers.

Vreuls, D., Obermayer, R.W.(, 1985. Human-system performance measurement in training simulators. Hum. Factors 27, 241–250.

Ward, M., Gruppen, L., Regehr, G., 2002. Measuring self-assessment: current state of the art. Adv. Health Sci. Educ. 7, 63–80.

Ward, M., Macrae, H., Schlachta, C., Mamazza, J., Poulin, E., Reznick, R., Regehr, G., 2003. Resident self-assessment of operative performance. Am. J. Surg. 185, 521–524.

Weber, D.E., Roth, W.-M., Mavin, T.J., Dekker, S.W., 2013. Should we pursue inter-rater reliability or diversity? An empirical study of pilot performance assessment. Aviation in Focus – Journal of Aeronautical Sciences 4, 34–58.

Weller, J., Shulruf, B., Torrie, J., Frengley, R., Boyd, M., Paul, A., Yee, B., Dzendrowskyj, P., 2013. Validation of a measurement tool for self-assessment of teamwork in intensive care. Br. J. Anaesth. 111 (3), 460–467.

West, B.T., Welch, K.B., Galecki, A.T., 2015. Linear Mixed Models: A Practical Guide Using Statistical Software. Second Edition. Chapman & Hall/CRC, New York.

Wieck, M.M., McLaughlin, C., Chang, T.P., Rake, A., Park, C., Lane, C., Burke, R.V., Young, L.C., Cleek, E.A., Morton, I., et al., 2018. Self-assessment of team performance using T-NOTECHS in simulated pediatric trauma resuscitation is not consistent with expert assessment. Am. J. Surg. 216, 630–635. https://doi.org/10.1016/j.amjsurg.2018.01.010.

Wildman, J.L., Salas, E., Scott, C.P., 2013. Measuring cognition in teams: a cross-domain review. Hum. Factors 56 (5), 911–941.

Xu, J., Anders, S.H., Pruttianan, A., France, D.J., Lau, N., Adams, J.A., Weinger, M.B., 2018. Human performance measures for the evaluation of process control human-system interfaces in high-fidelity simulations. Appl. Ergon. 73, 151–165.

Yule, S., Flin, R., Maran, N., Rowley, D., Youngson, G., Paterson-Brown, S., 2008. Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS behavior rating system. World J. Surg. 32 (4), 548–556. https://doi.org/10.1007/s00268-007-9320-z.