Empirical evaluation of the Process Overview Measure for assessing situation awareness in process plants

Nathan Lau, Greg A. Jamieson & Gyrd Skraaning Jr.

To cite this article: Nathan Lau, Greg A. Jamieson & Gyrd Skraaning Jr. (2016) Empirical evaluation of the Process Overview Measure for assessing situation awareness in process plants, Ergonomics, 59:3, 393-408, DOI: 10.1080/00140139.2015.1080310

To link to this article: http://dx.doi.org/10.1080/00140139.2015.1080310

Published online: 23 Sep 2015.

Submit your article to this journal

Article views: 122

View related articles

View Crossmark data
Empirical evaluation of the Process Overview Measure for assessing situation awareness in process plants

Nathan Lau, Greg A. Jamieson and Gyrd Skraaning Jr

Department of Industrial & Systems Engineering, Virginia Tech, Blacksburg, VA, USA; Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Canada; Industrial Psychology, OECD Halden Reactor Project, Halden, Norway

ABSTRACT

The Process Overview Measure is a query-based measure developed to assess operator situation awareness (SA) from monitoring process plants. A companion paper describes how the measure has been developed according to process plant properties and operator cognitive work. The Process Overview Measure demonstrated practicality, sensitivity, validity and reliability in two full-scope simulator experiments investigating dramatically different operational concepts. Practicality was assessed based on qualitative feedback of participants and researchers. The Process Overview Measure demonstrated sensitivity and validity by revealing significant effects of experimental manipulations that corroborated with other empirical results. The measure also demonstrated adequate inter-rater reliability and practicality for measuring SA in full-scope simulator settings based on data collected on process experts. Thus, full-scope simulator studies can employ the Process Overview Measure to reveal the impact of new control room technology and operational concepts on monitoring process plants.

Practitioner Summary: The Process Overview Measure is a query-based measure that demonstrated practicality, sensitivity, validity and reliability for assessing operator situation awareness (SA) from monitoring process plants in representative settings.

Introduction

Situation Awareness (SA) is an important dimension of human performance in process control. The notion, which refers to ‘knowing what is going’ (Endsley 1995b), is a criterion for both interface design (e.g. Burns et al. 2008; Tharanathan et al. 2010) and regulatory reviews (e.g. O’Hara et al. 2012). Thus, the most important application of the SA notion in process control is to the assessment of human performance.

A significant proportion of SA measurements are collected using the query/probe-based technique. First adopted by the Situation Awareness Global Assessment Technique (SAGAT; Endsley 1995a), the query/probe-based technique elicits declarative knowledge by directly questioning participants/workers in real time or during pauses in scenario trials/work settings. Hence, the connection between the measurements and the SA notion is intuitive compared to measurements based on task performance or physiological indicators. That is, the answers to the questions about the situation constitute (at least part of) the SA of the operators. In addition, the query/probe-based technique customises the queries or probes and administration methods according to the situation and awareness, thereby enhancing the coupling between the two to ensure construct validity of the SA measurements. The coupling between situation and awareness becomes particularly important in industrial settings where workers rely on their expertise, which is highly specific to their work domain. This contrasts with rating scales commonly used for other constructs (e.g. NASA TLX for workload) that rely on static sets of questions.

Development and evaluation of query/probe-based SA measures is limited in the process control domain. The Situation Awareness Control Room Inventory (SACRI; Hogg et al. 1995) is the only query-based measure developed for process control studies (Hallbert 1997; Sebok 2000; Lang et al. 2002). However, the methodological investigation and empirical assessment of SACRI is limited to its development work, which was based on a narrow selection of process control literature and relied primarily on pilot simulator studies with non-professional operators and one operator crew. Hence, there is a significant gap in methodological development pertaining to (i) the application of the query/probe-based technique to measure SA, and (ii) empirical assessment of the query/probe-based SA measures in process control. Further, recent review of the literature indicates a continual paucity of research in reliability and validity of similar evaluative methods in human factors (e.g. Stanton and Young 1999, 2003; Kanis 2014).

A companion article (Lau, Jamieson, and Skraaning in review) presents our domain-specific approach to advance...
the application of the query/probe-based technique to measure SA in process control. Process Overview is formulated to characterise SA acquired through monitoring process plants by synthesising field and observational studies. That is, Process Overview is a SA subdimension that represents the imprecise view of the plant conditions highlighting a subset of significant and relevant process behaviours. Process Overview is then employed as the basis to develop the methodological prescriptions in the Process Overview Measure. The Process Overview Measure assesses SA derived from monitoring process plants. The companion paper also includes a methodological comparison between SAGAT, SACRI and the Process Overview Measure, illustrating that adapting domain-general SA measures does not yield the same results as domain-specific ones (also see e.g. Durso and Drews 2010; Lau, Jamieson, and Skraaning 2012, 2013; Cooper, Porter, and Peach 2014).

This article reports on the empirical assessment of the Process Overview Measure in two full-scope nuclear power plant (NPP) simulator experiments. The results provide an indication of measurement properties – practicality, sensitivity, validity and reliability – when using the Process Overview Measure in realistic process control settings in which the coupling between situation and awareness is pertinent to operator performance. A holistic and questioning (as opposed to confirmatory) approach is adopted to discuss the current state of and future research needs for the Process Overview Measure. We then discuss empirical results of the Process Overview Measure that were consistent with the results of Scenario Understanding, task performance and debriefing data were considered validating evidence. Results that were inconsistent or conflicting among these data were deemed invalidating evidence of criterion validity. Validity was also evaluated by comparing whether the significant effects conformed to expectation or well-established hypotheses of the experimental manipulations. Validity was evaluated according to experimental effects revealed in the experiments. However, significant effects can be a misleading indicator of sensitivity because (i) experimental effects (or lack thereof) can be a result of effectiveness of experimental manipulations, rather than sensitivity of the measure; and (ii) experimental effects can be spurious, especially for large experimental designs. Because of these two threats, the validity of the experimental effects must also be assessed.

The Process Overview Measure must demonstrate sensitivity to reveal the effects of experimental manipulations. Sensitivity was evaluated by comparing significant effects of experimental manipulations and correlations between the Process Overview Measure and measurement on two other constructs. First, Scenario Understanding, another sub-SA dimension specific to process control, refers to mental representation and explanation of the control needs of the process plant (see, Lau, Jamieson, and Skraaning 2012). Scenario Understanding is formulated to characterise the knowledge or the SA acquired through reasoning and diagnosing control needs of the process plants (i.e. reasoning derived SA; Lau, Skraaning, and Jamieson 2015). Second, task performance refers to the overall effectiveness of control room operations demonstrated by the operators. Additionally, qualitative debriefing data from the participants were also employed for assessment of validity. Experimental results of the Process Overview Measure that were consistent with the results of Scenario Understanding, task performance and debriefing data were considered validating evidence. Results that were inconsistent or conflicting among these data were deemed invalidating evidence of criterion validity. Validity was also evaluated by comparing whether the significant effects conformed to expectation or well-established hypotheses of the experimental manipulations. This approach is not always available because human performance is difficult to predict for some experimental manipulations in full-scope simulator or industrial settings.

The Process Overview Measure must demonstrate reliability of the collected performance data. The initial evaluation of the Process Overview Measure focused on inter-rater reliability because the measure relied on process experts to determine the reference keys. This paper summarises the inter-rater reliability results; a full presentation of the evaluation method is reported by Lau, Jamieson, and Skraaning (2014).

Two full-scope simulator experiments provided opportunities to evaluate the Process Overview Measure against these

Methods and results of empirical assessment

This section first presents the overall evaluation approach and experimental environment, followed by the protocols and results of two full-scope NPP simulator experiments conducted to assess measurement practicality, sensitivity, validity and reliability of the Process Overview Measure in a representative process control environment.

General approach

The Process Overview Measure must demonstrate practicality in integrating the method into the preparation, data collection and analysis phases of high-fidelity simulator experiments. In addition, the recruited operators and process experts must be willing to accept the data collection procedure, given that congenial relationships with them are vital to data quality and future experiments. Practicality was judged subjectively, based on experience of the experimental research team and participants.
criteria. Further, the variation between testing environments was substantial. The first experiment focused on evaluating display types in a present digital control room environment. The second experiment focused on automation displays and staffing solutions in a futuristic control room environment. This variation provided a preliminary indication of the Process Overview Measure effectiveness across different process plant control rooms.

**General experimental environment**

Both experiments were performed in the HAlden Man-Machine LABoratory (HAMMLAB) on a high-fidelity simulator of a 1200 MW boiling water reactor (BWR) called HAMBO (Karls-son et al. 2001). The HAMMLAB simulator control room was equipped with multiple reconfigurable computer workstations for participating operators to monitor and control the simulated nuclear process. A large screen display in the middle of the control room provided an overview of the simulated plant. The exact set-up of the displays and workstations varied across the two experiments due to laboratory upgrades and experimental topics (see Figure 1). Interaction was carried out via keyboard and mouse. Operating procedures used in the experiments were standard printed procedures developed for the simulated plant. All participating operators had a personal telephone for communicating with plant maintenance and management. The experimental staff (e.g. researchers, technicians and process experts) managed the simulator, scenarios and data collection tasks (e.g. video and audio recording, expert ratings) from the observation gallery.

**Ethics review**

The Research Ethics Board at the University of Toronto reviewed and approved the protocols involving human participants for both experiments.

**Experiment 1 – present digital control room environment**

Experiment 1 investigated how operators used different types of displays in several scenarios for an operating environment
similar to control room settings of many existing modernised or newly built NPPs. The displays were developed for multi-monitor workstations that are becoming the standard replacements of hardwire panels. Burns et al. (2008) and Lau, Jamieson et al. (2008) presented the experimental method and discussed the interface design findings in detail. Experiment 1 provided empirical evidence on the Process Overview Measure’s relationships to Scenario Understanding and task performance, sensitivity to display and scenario manipulations in current control room settings, and validity on revealing experimental effects.

Method
Participants and operator roles. Six licensed operator crews (n = 6) were recruited from the BWR power plant on which the core and thermal hydraulic behaviours are modelled by the simulator. Each crew consisted of one reactor operator (RO) and one turbine operator (TO) 1. In two cases, participants currently working as ROs, but qualified as TOs, operated the turbine side. Because the displays were designed mainly to support operations of the turbine side, the measurements and data analysis of Experiment 1 focused on the TO (see Measures section).

Experimental manipulation
This study consisted of three experimental manipulations: display types, scenario types and scenario phases:

Three display types – Traditional, Advanced and Ecological – were compared. The Traditional displays were the computerised version of the hardwired wall panels, roughly representing the ‘state-of-practice’ design characterised by simplified versions of piping and instrumentation (P&ID) diagrams of the facilities with numeric outputs of instrumentation. The P&IDs also included equipment for operators to select and input control actions. The Advanced displays were a graphically enhanced version of the Traditional displays. The Advanced displays retained the simplified version of the piping and instrumentation diagrams in the Traditional displays and added some configural graphics (e.g. Bennett, Toms, and Woods 1993; Bennett, Nagy, and Flach 2006) and ‘mini-trends’ strategically developed and inserted by process experts. The Ecological displays (Lau, Veland, et al. 2008) were designed according to the Ecological Interface Design framework (Vicente and Rasmussen 1992). EID is a theoretical framework for designing human computer interfaces for complex systems. EID claims to enhance operator performance by specifying information requirements and perceptual features based on formative work analysis and cognitive controls, respectively.

The experiment included three Procedure-guided and three Knowledge-based scenarios. Procedure-guided scenarios were defined by a set of disturbances that could be resolved by referencing plant procedures. Scenarios in which disturbances could not be resolved by procedures were classified as Knowledge based. In other words, equipment failures anticipated by the utilities and job responsibilities familiar to operators characterised the Procedure-guided scenarios, while unanticipated and unfamiliar ones characterised the Knowledge-based scenarios. Skraaning et al. (2007) documented the details of each scenario.

Each scenario started with a ‘Detection’ phase, a time period just before the first alarm sounded, and ended with a ‘Mitigation’ phase that consisted of all subsequent events. (Figure 2 illustrates the detailed structure of the scenarios). The two phases afforded separate assessments of the effectiveness of the displays in supporting monitoring and intervention.

Experimental design
Each crew experienced all six combinations of the display-type and scenario-type manipulations (3 × 2) through six different scenarios (i.e. trials). The six experimental conditions (i.e. the combinations of display and scenario type) were counterbalanced using a Latin square technique. The scenario phase manipulation was not part of the counterbalancing as each crew experienced both phases in sequence for every scenario. The final experimental design was a 3 × 2 × 2 RBF (within-subjects) design with treatments of display type (Traditional, Advanced and Ecological), scenario type (Procedure guided and Knowledge based), and scenario phase (Detection and Mitigation). The treatments were completely crossed (N = 72 = 3 × 2 × 2 conditions × 6 crews).

Procedure
Each crew participated over three consecutive days. The first day consisted of initially obtaining informed consent and demographic information, followed by 6 h of training. The second day started with a 1-hour training session to refresh the materials presented on the first day, followed by three scenarios with 15-minute breaks in between. The third day started with three scenarios also with 15-minute breaks in between, followed by a debriefing/closing session.

For all trials, crews were asked to maintain the original power level and safe operation. A process expert registered scores to corresponding performance items at various points of the scenarios by observing the participants while they
monitored system states and resolved disturbances (see measure of task performance below). This process expert supported scenario development and served as a performance rater over 10 years at HRP; thus, the expert was a qualified rater for the experiment. The participants also responded to the Process Overview queries and subjective questionnaires during a short simulator freeze and at the end of each scenario. The simulation freeze occurred at the end of the Detection phase, which took up the first five to ten minutes of the scenario as depicted in Figure 2. The scenario then continued with the Mitigation phase, which was marked by the onset of the first alarm within the first minute. The Mitigation phase usually lasted about 30 min, followed by another administration of Process Overview queries and subjective questionnaires at the end of the scenario.

**Measures**

Measures of the Scenario Understanding and task performance results were employed to verify the sensitivity and validity of the Process Overview Measure. Lau, Skraaning, and Jamieson (2009), Lau, Jamieson, and Skraaning (2008) and Skraaning et al. (2007) reported the details of the performance measures and effects of the experimental manipulations.

The Process Overview Measure operationalized Process Overview as the accurate detection of meaningful changes in relevant process parameters (Lau, Jamieson, and Skraaning in review). In Experiment 1, the Process Overview Measure included both context- and fault-sensitive queries about the parameter changes in the recent past. The experimenter administered the queries to RO and TO individually during one simulator freeze and at the end of every trial (Figure 2). For this study, only the TO data was analyzed given the experimental displays were designed for the turbine side. Fault-sensitive queries administered during the freeze could produce cuing effects, which could alter the operator behaviours in the Mitigation phase (i.e., second phase of the scenario). To standardize the behavioural changes due to queries across operators, the Mitigation phase began with an alarm that alerted the operators to faults developing during the Detection phase (i.e., first phase of the scenario).

Scenario understanding was measured by the Halden Open Probe Elicitation (HOPE) measure (Skraaning et al. 2007). HOPE employed process experts to rate the operator’s understanding of the scenario on a four-point-anchored scale. The process experts performed their ratings based on (i) observation (e.g. conversation between operating crew members) from the experimental gallery, and (ii) telephone inquiries to operators about the plant states. Process experts or experimental leaders, disguised as plant management, telephoned operators to inquire about plant states at strategically predefined times of the scenarios. HOPE prescribed a list of ‘open’ probes, which did not refer to any scenario-specific characteristics, to avoid cuing effects. For example, a probe can be: ‘Is everything normal?’ During the telephone calls, the simulator was frozen but the operators could access all process displays. For this experiment, a process expert conducted two ratings on the TOs – one at the middle and the other at the end of the scenario phase. In contrast, the Process Overview data was collected only during the simulator freeze. The final score for each phase was the average of the two ratings.

Task performance was measured by the Operator Performance Assessment System (OPAS; Skraaning 1998, 2003; Skraaning et al. 2007). OPAS provided a structure for the assessment of whether operators carry out their task work in accordance with scenario solutions specified a priori by process experts. Prior to data collection, process experts analyzed the scenarios and developed optimal solutions by identifying items capturing the desired performance. During the experiment, a process expert registered the points earned by operators or crews in completing the predefined activities within each performance item based on observations of operator verbalization, physical behaviours, and system states. The employed performance index was the unweighted average of all performance items defined for a scenario. The OPAS index reflected the degree of conformance between operator/crew performance and predefined optimal solutions to scenarios. For this experiment, the OPAS indicated task performance of the crew as points were awarded whether the RO or TO completed the predefined activities.

**Results**

The Process Overview Measure did not correlate with HOPE ($r(72) = -0.34, p = .78$) or OPAS ($r(72) = -0.15, p = .90$).

The Process Overview Measure data were analyzed with ANOVA using Type III/Unique sums of squares with fixed factors of display type (Traditional, Ecological and Advanced), scenario type (Procedure-guided and Knowledge-based), and scenario phase (Detection and Mitigation) that were fully crossed, and a random factor of crew. The significant effects on the Process Overview Measure are the main effect of scenario phase ($F(1, 55) = 8.28, p = .01, \eta^2 = .13$), and the three-way interaction effect of display type, scenario type and scenario phase ($F(2, 55) = 2.93, p = .06, \eta^2 = .10$; Figure 3). However, Levene’s test only indicated homogeneity of variance for the three-way interaction effect ($F(11, 60) = 1.44, p = .18$) but heterogeneity of variance for scenario phase ($F(1, 70) = 3.30, p = .07$). A non-parametric, Wilcoxon-Matched Pairs test on scenario phase ($U(36) = 153, z = 2.47, p = .01$), which was robust to deviation from homogeneity of variance, confirmed the ANOVA scenario phase effect.

The HOPE data were analysed with a multivariate ANOVA using Type III/Unique sums of squares with fixed factors of display type (Traditional, Ecological and Advanced), scenario type (Procedure guided and Knowledge based) and scenario phase (Detection and Mitigation). The significant effects on
the HOPE measurements are the main effect of scenario phase \(F(1, 5) = 20.38, p = .01\), the two-way interaction effect of display type and scenario type \(F(2, 4) = 9.98, p = .03\), and the three-way interaction effect of display type, scenario type and scenario phase \(F(2, 4) = 31.07, p = .00\); Figure 4).

OPAS data were analysed in an ANCOVA with fixed factors of display type, scenario type and scenario phase, a random factor of crew, and a covariate of Workload. The ANCOVA was an over-parameterised model built on Type II sums of squares. This analysis explored the fixed effects on OPAS measurements controlled for Workload, assessing the support for problem-solving provided by each display type while limiting the mediating effect of task demand. The significant effects on OPAS after controlling for Workload are the two-way interaction of display and scenario \(F(2, 10) = 8.09, p < .01\), and the three-way interaction of display, scenario and phase \(F(2, 9) = 6.08, p = .02\); Figure 5). A post hoc analysis following the technique proposed by Howell (2002) and using Tukey’s Honestly Significant Difference (HSD) criterion for significance was conducted to confirm the performance advantage of Ecological displays in the Detection phase of Knowledge-based scenarios. The final Tukey’s HSD post hoc analysis indicated that the performance of Ecological displays \((M = 1.81, SD = .89)\) was significantly higher than Traditional \((M = .09, SD = .72, p < .01)\) and Advanced \((M = .28, SD = .87, p < .01)\) in the Detection phase of Knowledge-based scenarios.

**Summary**

The Process Overview Measure did not correlate with HOPE or OPAS measurements. This finding is slightly surprising as weak-to-moderate correlations were expected as validating evidence that these measures are all distinct indicators of human performance, although SA studies in the literature do not always observe correlation with other performance indicators (e.g. O’Brien and O’Hare 2007). On the other hand, HOPE and OPAS measurements were strongly correlated, suggesting a strong overlap between the measurements. Further empirical evidence would be necessary to study the convergent and discriminant validity of these measurements.

The Process Overview Measure revealed several experimental effects that (i) supported the expected impact of the experimental manipulations and (ii) corroborated the significant effects on HOPE and OPAS. Consistent with interface design research on EID (e.g. Vicente 2002; Burns and Hajdukiewicz 2004), the results indicated that Ecological displays facilitated Process Overview better than Traditional and Advanced displays in the Detection phase of Knowledge-based scenarios. Consistent with other measurements, Process Overview demonstrated a similar graphical pattern to Scenario Understanding and Task Performance. The interpretation of the converging results of the three measures must account for the lack of correlation between the Process Overview Measure and either HOPE and OPAS, but the strong correlation between HOPE and OPAS. In other words, the experimental effects of Process Overview Measure converge with two other but potentially redundant measures.
operators to manage the faults and to resume automation. However, the automation could be obstructed by process start-up or shutdown procedures of the simulated plant. Process deviations in normal plant states, and executed was implemented as computer scripts that monitored basic control room settings in future NPPs. The advanced automation was implemented as computer scripts that monitored basic control room settings in future NPPs. The advanced automation in managing a NPP, given different automation displays and staffing configurations. The advanced automation and large screen display created an operating environment that could be representative of control room settings in future NPPs. The advanced automation was implemented as computer scripts that monitored basic process deviations in normal plant states, and executed start-up or shutdown procedures of the simulated plant. However, the automation could be obstructed by process faults throughout the scenarios, and would ‘pause’ for the operators to manage the faults and to resume automation. Such automation is expected to be available in the future. For assessing measurement properties, Experiment 2 provided further empirical evidence on the Process Overview Measure’s sensitivity and validity to experimental manipulations, and generalisability of use in different control room settings.

**Experiment 2 – futuristic control room environment**

Experiment 2 investigated how operators cooperated with advanced automation in managing a NPP, given different automation displays and staffing configurations. The advanced automation and large screen display created an operating environment that could be representative of control room settings in future NPPs. The advanced automation was implemented as computer scripts that monitored basic process deviations in normal plant states, and executed start-up or shutdown procedures of the simulated plant. However, the automation could be obstructed by process faults throughout the scenarios, and would ‘pause’ for the operators to manage the faults and to resume automation. Such automation is expected to be available in the future. For assessing measurement properties, Experiment 2 provided further empirical evidence on the Process Overview Measure’s sensitivity and validity to experimental manipulations, and generalisability of use in different control room settings.

**Method**

**Participants and operator roles.** Nine licensed operator crews (n = 9) were recruited from three Swedish NPPs. Plants 1, 2 and 3 provided four, two and three crews, respectively. Plants 1 and 3 operated with a similar nuclear process to the HAMMLAB simulator, whereas Plant 2 operated with an identical nuclear process. Each three-person crew consisted of one RO, one TO and one shift supervisor employed at the same plant. Some participants within each crew worked together on a daily basis while others did not.

**Experimental manipulations**

Experiment 2 consisted of three experimental manipulations – staffing solution, automation interface and scenario period. It also included a quasi-experiment independent variable of previous plant experience, which was an outcome of the recruiting process.

Two types of staffing solution – Traditional and Untraditional – were included to explore new staffing complements facilitated by high levels of plant automation. The Traditional staffing solution consisted of a RO, TO and Shift Supervisor according to the conventional composition of an operating crew in their own and most NPPs. The Traditional staffing solution was responsible for operating one simulated, highly automated NPP. The RO and TO handled the reactor and the turbine sides of the nuclear process, respectively, while the SS managed and supported the RO and TO. Participants were assigned to roles according to their work positions at their home plant. The Untraditional staffing solution consisted of a Main Operator (MO), Assistant Operator (AO) and Work Manager (WM). This three-person staff composition was formulated to explore staffing considerations in a hypothetical future NPP. The Untraditional staffing solution was responsible for operating two highly automated NPPs in two separate control rooms, unlike the Traditional staffing operating a single plant at the normal crew size. The MO was responsible for both the reactor and turbine side of simulator Plant A in HAMMLAB, while the AO was responsible for the turbine side of simulator Plant B in an adjacent room. The WM was responsible for both simulator Plants A and B, managing and supporting both MO and AO. The WM was also responsible for reallocating the AO from Plant B to Plant A to support MO under critical operating conditions. Plant B operated in ‘pause’ mode whenever AO was relocated to Plant A. When AO returned, Plant B resumed operations. Eittruehm et al. (2010) present the details of this experimental manipulation.

Two types of automation interface – Non-transparent and Transparent – were included to investigate the effects of displaying automation information in supporting operators to cope with high levels of plant automation. The Non-transparent and Transparent automation interfaces shared the same interactive functions but differed in the amount of information presented about the plant automation. The Non-transparent automation interface presented minimal information about plant automation containing only (i) the interaction buttons and (ii) indication of the unsatisfied condition that caused plant automation to stop. Operators had to make inferences about automation activity from observing process events.
and changes in process parameters based on the expected sequence of actions as prescribed by the operating procedures. The Transparent automation interface consisted of a plant automation overview and detailed displays presented on two 30” LCD monitors, respectively. The overview display showed the goals and effects of process control tasks and the interaction buttons to control the plant automation. The detailed automation display showed historical and ongoing automation activities through excerpts from mimic displays and a chronological log. (Hurlen et al. (2010) and Skraaning et al. (2010) document the details of this experimental manipulation.)

Each scenario started with an Easy followed by a Difficult scenario period. The Easy scenario period contained process faults that were independent of each other. This scenario period was assumed to have low time pressure and reliable automation (i.e. operating within the design basis of the plant automation). The Difficult scenario period contained more complex faults that could be either dependent or independent of each other. The causes of the faults were difficult to understand and time pressure was typically high, particularly if unfinished tasks from the Easy period had to be completed in parallel with handling the complex faults introduced in this period. Figure 6 illustrates the structure of the scenario.

The participating operators were recruited from three different operating NPPs – Plants 1, 2 and 3. Plants 1 and 3 operate with a reactor core and thermal process similar to those modelled in the HAMBO simulator, whereas Plant 2 operates with a reactor core and thermal process identical to those modelled in HAMBO. Therefore, Plant 2 operators were expected to have more control experience on the HAMBO simulator than Plants 1 and 3 operators.

**Experimental design**

Each three-person crew experienced all four combinations of the staffing solution and automation interface manipulations (2 × 2) twice through in eight different scenarios (i.e. trials). The four experimental conditions (i.e. the combinations of staffing solution and automation interface) were randomly assigned to the eight scenarios and randomly presented to each crew. The scenario period manipulation was not part of the randomisation as each crew experienced both periods in sequence for every scenario.

Because the plant factor was mainly an outcome of the recruiting process (i.e. a quasi-experiment independent variable) rather than deliberate experimental manipulations, the plant factor was examined with its own model.

For assessing sensitivity of the Process Overview Measure with respect to experimental manipulations, the experimental design was a 2 × 2 × 2 RBF (within-subjects) design with treatments of staffing solution (Traditional and Untraditional), automation interface (Non-transparent and Transparent) and scenario period (Easy and Difficult). The treatments were completely crossed (N = 72 = 2 × 2 × 2 conditions × 9 crews). For assessing sensitivity with respect to the experience of the participants with the simulated process, a one-way ANOVA was employed with the treatment of plants (1, 2 and 3).

**Measures**

The Process Overview Measure was used to assess monitoring-derived SA (i.e. Process Overview). During one simulator freeze and at the end of every trial, the queries were administered to RO and TO individually in the Traditional and MO in the Untraditional staffing conditions.

Experiment 2 only included context-sensitive parameters, omitting fault-sensitive queries. There were three reasons for this choice of implementation. First, the Untraditional staffing solution required the MO to assume the responsibilities of both reactor and turbine sides; consequently including the fault-sensitive queries would have been too time-consuming, interfering with the operators in the MO role. Second, the scenarios were designed to introduce faults in increasing order of difficulty in anticipation that participants performing well could proceed further in the scenarios than those performing poorly. This prohibited effective use of fault-sensitive parameters for querying and alarms for standardising the cuing effects. As mentioned in the companion article, fault-sensitive queries are specific to the process faults in the scenarios and thus introduce significant cuing. Given that the operators/crews would likely progress at different pace and thus experience different faults in this experiment, selecting fault-sensitive parameters became impractical. Further, the number of alarms at the start of the Difficult period could not be controlled as the operators/crews might have carried over alarms as well as unfinished tasks from the Easy to Difficult periods. Third, the threat of ceiling effects of relying on context-sensitive parameters was reduced as many of the operators were working on a simulated plant that differed from the plant they normally operated. That is, operators might be sampling context-sensitive parameters at their home plants very effectively, given their experience and training on the plant behaviours and displays; hence, the Process Overview Measure could have a ceiling effect. However, their experience and training would be less effective for monitoring another plant. The omission of fault-sensitive queries impacted content validity of the Process Overview Measure in Experiment 2. Thus, the interpretation of the Process Overview results must take this implementation choice into account.

Scenario Understanding was measured by the Automation and Scenario Understanding Rating Scales (ASURS; Lau 2012). ASURS employs process experts to rate operator understanding of automation and the scenario using four items. The first item assesses the operator awareness of automation purpose (for the scenario) on a four-point anchored scale. The second and third items assess operator awareness of process (or actions) and performance of automation on a
five-point anchored scale, respectively. The final item assesses operator understanding of the overall situation on a five-point anchored scale. During data collection, process experts perform their ratings based on observation (e.g. of conversation between operating crew members) from the experimental gallery. The final Scenario Understanding index was the sum of all the ratings. In Experiment 2, a process expert conducted the ratings of the four items for each scenario period. Unlike Experiment 1 in which HOPE averaged 2 ratings, the ASURS ratings occurred once per scenario period that coincides with data collection of the Process Overview Measure. The process experts rated the combined understanding of the RO and TO in the Traditional staffing conditions, but just the MO in the Untraditional staffing conditions. The trials of the Plant 1 crews were used for training the process expert on ASURS; thus, the corresponding data were omitted in the analysis.

Task performance was measured by OPAS as in Experiment 1. OPAS indicated task performance of the crew as scores were awarded irrespective of which operators completed the predefined activities. In Experiment 2, the OPAS items were separated into Detection and Operation categories to isolate specific aspects of task performance. By comparison to Experiment 1, the OPAS items were formulated to be more concrete with fewer references to cognitive activities such as problem-solving activities and strategies to improve discriminant validity with other measures. Specifically, reducing references to cognitive activities was expected to lower the correlations between measures of Scenario Understanding and task performance.

**Procedure**

Each crew participated over 4 days. The first day and the morning of the second day were dedicated to obtaining informed consent and demographic information; introducing the study; and training the participants.

The data collection began after lunch on the second day and continued until lunch on the fourth day. Each crew participated in eight scenarios related to starting up the highly automated simulator plant. At the beginning of each scenario, a process expert briefly explained the general context of the scenario. Eitrheim et al. (2010) and Skraaning et al. (2010) reported details of the scenarios. Each trial began with the Easy scenario period of about 15 minutes followed by a simulator freeze to collect responses to the Process Overview queries and questionnaires. After the simulator froze, the trial entered the Difficult scenario period of about another 20 minutes. At the end of the Difficult scenario period, operators responded to another set of Process Overview queries and questionnaires. Figure 6 illustrates the general structure of the scenario.

For trials of two crews, other process experts were recruited to conduct performance ratings of the operators to assess inter-rater reliability of several human performance measures (Lau, Jamieson, and Skraaning 2014).

To minimise overload and carry-over effects, a 20-minute break separated every scenario. The data collection concluded with a 90-min debriefing interview after lunch on the fourth day to obtain feedback on the experiment, especially the experimental manipulations.

**Analysis and results**

**Process Overview.** The Process Overview Measure did not correlate with ASURS (r(80) = .138, p = .24) or OPAS-Detection (r(144) = .07, p = .43), but did correlate slightly with OPAS-Operation (r(144) = .18, p = .03). The slight positive correlation between Process Overview and OPAS-Operation occurred in the expected direction but the shared variance (r² = .03) was extremely low.

To test the influence of control experience on the HAMBO simulator process, the Process Overview data were analysed in a one-way ANOVA using Type III/Unique sums of squares with a fixed, between-subjects factor of Plant (1, 2 and 3). The one-way ANOVA revealed a difference in Process Overview between plants where operators work (F(2, 69) = 2.90, p = .06, η² = .08; Figure 7). Levene's test indicated homogeneity of variance for plant (F(2, 69) = .91, p = .41).

To assess sensitivity to experimental manipulations, the Process Overview Measure was analysed with ANOVA using Type III/Unique sums of squares with fixed factors of scenario period (Easy and Difficult); staffing solution (Traditional and Untraditional); and automation display (Non-transparent and Transparent) that were fully crossed, and a random factor of crew. The Process Overview Measure revealed the main effect of staffing solution (F(1, 56) = 2.84, p = .09, η² = .05); and the interaction effect of staffing solution and automation display (F(1, 56) = 3.70, p = .05, η² = .07; Figure 8). Levene’s tests indicated homogeneity of variance for the two-way interaction of staffing solution and automation display (F(3, 68) = .91, p = .44) but slight heterogeneity of variance for staffing solution (F(1, 70) = 2.80, p = .10). In addition, there appeared to
be a slight correlation between means and standard deviation for the effect of staffing solution. A non-parametric, Wilcoxon-Matched Pairs test on staffing solution ($U(36) = 210.5, z = 1.71, p = .09$) provided some statistical confirmation on the ANOVA staffing solution main effect (indicating the robustness of ANOVA to deviations in homogeneity of variance).

**Corroborating evidence**

ASuRS measurements correlated moderately with both OPAS-Detection ($r(80) = .47, p = .00$) and OPAS-Operation ($r(80) = .47, p = .00$). Note that these correlations must be interpreted with care because of the difference in the aggregation of the ASuRS and OPAS scores. The aggregation of ASuRS scores excluded data collected from the roles of Shift-supervisor and WM in the Traditional and Untraditional staffing solution, respectively. On the other hand, the aggregation of both OPAS scores included data collected from the roles of Shift-supervisor and WM (i.e. the full crew). OPAS-Detection and OPAS-Operation measurements demonstrated a moderate correlation ($r(144) = .62, p = .00$).

To test the influence of process knowledge about the simulator based on the NPPs where operators work, ASuRS, OPAS-Detection and OPAS-Operations were analysed in an one-way ANOVA using Type III/Unique sums of squares with a fixed, between factor of Plants (2 and 3) and a random factor of crew. The significant effects were present in all three measures – ASuRS ($F(1, 38) = 2.76, p = .05, \eta^2 = .09$, Figure 9); OPAS-Detection ($F(2, 69) = 2.76, p = .07, \eta^2 = .07$, Figure 10); and OPAS-Operation ($F(2, 69) = 5.66, p = .01, \eta^2 = .14$, Figure 10).

The ASuRS, OPAS-Detection and OPAS-Operation data were analysed in an ANOVA using Type III/Unique sums of squares with fixed factors of scenario period (Easy and Difficult), staffing solution (Traditional and Untraditional), automation display (Non-transparent and Transparent) and a random factor of crew.

ASuRS revealed a main effect of scenario period ($F(1, 4) = 50.54, p < .01$), which was not related to the manipulations that had effects on Process Overview. Due to the lack of overlap of experimental effects, Scenario Understanding results could not provide any additional evidence corroborating the results from the Process Overview Measure.

OPAS-Detection revealed main effects of period ($F(1, 8) = 30.10, p = .00, \eta^2 = .79$); staffing solution ($F(1, 8) = 17.31, p = .00, \eta^2 = .68$); automation display ($F(1, 8) = 4.42, p = .07, \eta^2 = .36$); and the interaction effect of scenario period and staffing solution ($F(1, 8) = 3.44, p = .1$). The interaction effect of scenario period and staffing solution (Figure 11 left) was particularly relevant to verifying the validity of the staffing solution and automation display interaction effect revealed by the Process Overview Measure. OPAS-Operation revealed main effects of period ($F(1, 8) = 14.73, p = .00, \eta^2 = .65$); staffing solution ($F(1, 8) = 5.38, p = .05, \eta^2 = .40$); and the interaction effect of scenario period and staffing solution ($F(1, 8) = 3.74, p = .09, \eta^2 = .32$). The interaction effect of scenario period and staffing solution (Figure 11 right) was particularly relevant to verifying the validity of the staffing solution and automation display interaction effect revealed by the Process Overview Measure.
Experiment 1. That is, the Process Overview Measure appeared discriminant from ASuRS and OPAS, especially in comparison to the correlation between ASuRS and OPAS. The Process Overview Measure revealed several experimental effects that supported the expected impact of experimental manipulations but only partially corroborated with those of ASuRS and OPAS. The strongest evidence of sensitivity and validity of the Process Overview Measure was the significant effect of the quasi-experimental factor of plant. The effect supported the expected outcome that operator experience with the plant process improved SA and task performance. In addition, the plant effect on the Process Overview Measure was fully consistent with the ASuRS and OPAS (Detection and Operation). On the other hand, the Process Overview Measure illustrated that performance was best in the untraditional staffing solution with Transparent automation displays in a two-way interaction effect. This effect was in general conflict with the OPAS-Detection two-way interaction effects illustrating that the untraditional staffing solution led to worse performance than the Traditional staffing solution, especially in the Difficult scenario periods. Based on the debriefing, a plausible explanation was that the operators were fixated on monitoring the automation given ‘useful’ displays when the staff level was lower than the conventional arrangement. This inappropriate allocation of attention compromised the operators in identifying process faults and taking control actions.

The Process Overview Measure demonstrated fair-to-substantial inter-rater reliability. Debriefing data suggested that loss of agreement could be due to variability in interpreting the term ‘recently’ to answer the Process Overview queries. ‘Recently’ became ambiguous when operators thought of multiple events during the scenario that could be considered important for defining the relevant time period to judge parameter changes. In addition, some parameters fluctuated during the scenarios, challenging the operator to select a response in the set without some predefined time periods in the query.

Summary
The Process Overview Measure did not correlate with ASuRS and OPAS measurements, while ASuRS and OPAS correlated moderately, generally consistent with the results in Experiment 1. That is, the Process Overview Measure appeared discriminant from ASuRS and OPAS, especially in comparison to the correlation between ASuRS and OPAS.

The Process Overview Measure revealed several experimental effects that supported the expected impact of experimental manipulations but only partially corroborated with those of ASuRS and OPAS. The strongest evidence of sensitivity and validity of the Process Overview Measure was the significant effect of the quasi-experimental factor of plant. The effect supported the expected outcome that operator experience with the plant process improved SA and task performance. In addition, the plant effect on the Process Overview Measure was fully consistent with the ASuRS and OPAS (Detection and Operation). On the other hand, the Process Overview Measure illustrated that performance was best in the untraditional staffing solution with Transparent automation displays in a two-way interaction effect. This effect was in general conflict with the OPAS-Detection and Operations two-way interaction effects illustrating that the untraditional staffing solution led to worse performance than the Traditional staffing solution, especially in the Difficult scenario periods. Based on the debriefing, a plausible explanation was that the operators were fixated on monitoring the automation given ‘useful’ displays when the staff level was lower than the conventional arrangement. This inappropriate allocation of attention compromised the operators in identifying process faults and taking control actions.

The Process Overview Measure demonstrated fair-to-substantial inter-rater reliability. Debriefing data suggested that loss of agreement could be due to variability in interpreting the term ‘recently’ in the queries. The inter-rater reliability should be acceptable for use in representative settings but improvements on the queries would be desirable. No other operator and process expert feedback indicated that the Process Overview Measure was difficult to implement or collect data for in Experiment 2, suggesting its practicality for...
representative process control settings. However, fault-sensitive parameters were omitted to suit the design of the scenario structure and the Untraditional staffing. This decision did simplify the Experiment 2 implementation of the Process Overview Measure.

Discussion

The Process Overview Measure was developed as a methodological progression towards a query-based SA measure specific to the process control domain, particularly for assessing knowledge gained from monitoring of process plants. The Process Overview Measure applies insights from field research to specify methodological details such that variability can be minimised across implementations in representative process control experiments (Lau, Jamieson, and Skraaning in review).

The Process Overview Measure was evaluated for practicality, sensitivity, validity and reliability in two full-scope simulator experiments. Practicality was assessed by the qualitative comments of the participants and process experts, who did not indicate any major challenges in incorporating the method into the preparation and data collection for the two representative experiments. However, some operators in Experiment 2 indicated that the orienting term ‘recently’ needed to be clarified. A follow-on study explored alternative query structures with the goal of improving inter-rater reliability of the Process Overview Measure (Lau et al. 2011). In summary, the Process Overview Measure is deemed practical for assessing SA in representative process control settings.

The sensitivity of the Process Overview Measure was assessed according to experimental effects and corroborating evidence provided by measures of Scenario Understanding and task performance. For Experiment 1, the experimental effects of the Process Overview Measure corroborated with ones of HOPE and OPAS, providing a strong indication of sensitivity.

The sensitivity assessment based on experimental effects was more complex in Experiment 2. The effects of plant factor in the one-way ANOVAs revealed by the Process Overview Measure, ASuRS, OPAS-Detection and Operation, all converged, indicating validity of the experimental effects. The plant factor provided a particularly valid test because the expected results were not related to the experimental manipulations developed by researchers. However, the Process Overview Measure revealed some opposite effects of the experimental manipulations from those of OPAS-Detection and Operation. While different measures were expected to reveal different experimental effects, the opposite indication of experimental manipulations between Process Overview and task performance was unexpected, especially given the results on the plant factor and Experiment 1. In essence, the significant effects of the experimental manipulation revealed by the Process Overview Measure indicated sensitivity without corroborating evidence from other measures.

The Experiment 2 debriefing provided a plausible explanation of this result. The operators stated that that they relied heavily on the Transparent automation displays and felt understaffed with the Untraditional staffing solution when multiple faults were occurring. The interaction effect of Process Overview could be a consequence of operator tendency to fixate on monitoring with the Transparent automation interface when they were overwhelmed by scenario challenges while operating at the lower than conventional staffing level. The debriefing data provided another indication that the interaction effect revealed by Process Overview was not a result of chance but of experimental manipulations. If the corroborating evidence of the plant effect and debriefing comments were absent, the experimental effects on Process Overview should be considered spurious. Taking all the evidence together in Experiments 1 and 2, the Process Overview Measure demonstrated sufficient sensitivity in revealing experimental effects in representative process control experiments.

The Process Overview Measure demonstrated moderate inter-rater reliability according to \( \kappa \) results based on ratings between three process experts (Lau, Jamieson, and Skraaning 2014). For the complexity of the judging parameter behaviours of a complex process, the Process Overview Measure is sufficiently reliable for assessing monitoring performance in representative process control settings.

The Process Overview Measure did not correlate with measures of Scenario Understanding (i.e. HOPE and AUSRS) or measures of task performance (i.e. OPAS) in either experiment, except for a significant but negligible correlation with OPAS-Operation in Experiment 2. In contrast, HOPE and ASuRS correlated with OPAS in both experiments. The lack of correlation for Process Overview Measure with measures of Scenario Understanding is slightly surprising as weak-to-moderate correlations were expected as validating evidence that these measures are distinct indicators of SA. That is, there should be sufficient correlation to indicate that the Process Overview Measure is related but not identical to measures of Scenario Understanding. Similarly, a weak correlation between the Process Overview Measure and OPAS should be present as effective monitoring as it is believed to be necessary for effective control of process plants to indicate criterion validity. The investigation of the sensitivity indicated that the experimental effects revealed the Process Overview Measure should be valid. Thus, these empirical results only permit inferences that monitoring tasks appear to be substantially different from both reasoning and control tasks, and criterion validity cannot be established with correlation statistics in two representative experiments, which might not have the necessary range of scenarios.
The literature also contains empirical results on the lack of correlations between SA subdimensions and task performance (e.g. Lang et al. 2002; Jones and Endsley 2004; O’Brien and O’Hare 2007). In fact, no SA measures appear to have firmly established criterion validity in representative industrial settings. Two factors appear to contribute to this finding in our studies and the literature. First, correlation between measures can be difficult to reveal with small sets of sampled scenarios, given complex operator performance behaviours. For instance, Lang et al. (2002) were able to find a significant but weak correlation between SACRI and their Scenario Understanding measure ($r(80) = .24$; 6% shared variance). However, a small effect is very difficult to replicate when there are substantial differences between studies. In the Lang et al. study, SACRI measurements did not correlate with their task performance/criterion measurements as in the studies reported here. A full range of scenarios may help establish criterion validity for many query/probe-based SA measures.

Second, the Process Overview queries might not cover enough content of Process Overview (monitoring-derived SA) to reveal associations with other subdimensions and task performance. That is, the Process Overview Measure might not be sufficient to capture the empirical connection between Process Overview (i.e. monitoring-derived SA) and Scenario Understanding (i.e. reasoning-derived SA) as well as task performance (see, Lau and Skraaning 2015). The lack of correlation between SA subdimensions is not unique to these studies. Correlations between three levels of SA are not always present in SAGAT studies (e.g. Jones and Endsley 2004; O’Brien and O’Hare 2007). The evidence in this research as well as the literature suggests that the empirical relationships between performance of monitoring, reasoning and control executions most likely need to be modelled with additional moderating factors using more sophisticated techniques.

### State of current and future research

The Process Overview Measure is the most theoretically and empirically investigated query/probe-based SA measure for process control, establishing a foundation for researchers and practitioners to apply the measure and interpret the measurements. However, the current empirical work on Process Overview Measure cannot represent full validation. Kanis (2014) questions the concept of (permanent) validation for measures involving human agency and proposes a questioning – as opposed to a confirmatory – approach for studying the properties of measurements. That is, measurement evaluation studies are conducted to illustrate the best ways to apply, interpret and improve the measures. Thus, the discussion of the Process Overview Measure concludes with the state of current and future research following the assessment outlines prescribed by Stanton (2014).

Table 1 summarises the current research of the Process Overview Measure.

Table 1 indicates a need for additional empirical studies for validity and reliability assessment of the Process Overview Measure. Empirical investigation should first address operator feedback regarding clarifying the queries and response options. These modifications are expected to improve inter-rater reliability. Three temporary solutions may improve inter-rater reliability. First, overall reliability can improve with aggregating data from multiple raters (Kraemer 1992), although recruiting several experts can be difficult and costly. Second, the expert(s) can review simulator logs to improve
consistency, but this may lead to trade-off with construct validity. That is, expert judgement of parameter changes with log reviews is no longer in situ, so there is a risk of converting a semantic processing task into visual matching one. This trade-off deserves empirical investigation. Finally, the experts can discuss judgement criteria prior to experimental trials to standardise judgement but the inter-changeability of the raters would then be uncertain. Lau, Jamieson, and Skraaning (2014) discuss the inter-rater reliability results and challenges of the Process Overview Measure in detail.

Criterion and discriminant validity can be further assessed through meta-analysis with data from additional studies employing the same measures. Parallel form reliability can be assessed through similarities in operator scores between different sets of parameters identified by different process experts. Concurrent validity and internal consistency will require development of another SA measure of monitoring process plant and new statistical analysis method, respectively.

Conclusion

The Process Overview Measure is a query-based SA measure that operationalises Process Overview, a domain-specific SA characterisation depicting operator knowledge acquired through monitoring process plants (see companion article). By adopting a domain-specific approach, the Process Overview Measure prescribes unique methodological details to maximise the coupling between situation and awareness in the SA measurements and standardise the queries and administrative methods across studies. The Process Overview Measure demonstrates practicality, sensitivity, validity and reliability as a SA measurement tool in two full-scope simulator experiments, establishing an initial empirical basis to guide researchers and practitioners on the application of the measure and interpretation of the results. The empirical basis for Process Overview Measure is not only essential for SA measurements in the process control domain, but also relevant to understanding the effectiveness of the query/probe-based technique across SA measures and domains. Future work needs to address the potential improvements in inter-rater reliability based on operator feedback on the queries and response options. Additional empirical studies are also necessary to provide further evidence and support meta-analysis to validate the Process Overview Measure. This research initiates the process of establishing a theoretical and empirical foundation for the Process Overview Measure, a query/probe-based SA measure to assess how operators monitor process plants. This foundation suggests that the Process Overview Measure should be considered a practical choice for assessing monitoring derived SA in validation and verification activities of human performance or control room designs.

Notes

1. At the BWR plant where operators were recruited, each crew consisted of a shift supervisor as well as reactor and TOs.
2. In this experiment, automation operated the reactor side of Plant B to start up the plant with the AO.
3. Note that the aggregation of Process Overview excludes data collected from the role of Shift-supervisor and WM in the Traditional and Untraditional staffing solution, respectively, because ASURS was intended to measure the Scenario Understanding of the operators only. Further, ASURS data from the first four crews (Plant 1) were omitted from the analysis because the process experts (i.e. raters) were retrained on the measurement protocol.
4. The Mann-Whitney U-test yielded a p-value of .09, which confirmed the main effect of staffing solution. However, both the parametric and non-parametric tests yielded p-values between .05 and .1, which were weak indications of experimental effect.
5. Note that correlations between measures should generally not be very strong (i.e. r^2 > .5) as that is often an indication that the measures are only superficially different, measuring the same underlying construct (i.e. lack of discriminant properties).

Disclosure statement

No potential conflict of interest was reported by the authors.

References


Hogg, David N., Knut Follesø, Frode Strand-Volden, and Belén Hallbert, Bruce P. 1997. “Situation Awareness and Operator


