The Process Overview Measure: Methodological Developments to Enhance Inter-Rater Reliability

Nathan Lau, Gyrd Skraaning jr, Tommy Karlsson, Christer Nihlwing, & Greg A. Jamieson

CEL 11-02

Cognitive Engineering Laboratory, Department of Mechanical & Industrial Engineering, University of Toronto
5 King's College Road, Toronto, Ontario, Canada M5S 3G8
URL: www.mie.utoronto.ca/labs/cel/

Directors: Kim J. Vicente, Ph.D., P. Eng.
Greg A. Jamieson, Ph D., P. Eng.

The Cognitive Engineering Laboratory (CEL) at the University of Toronto (U of T) is located in the Department of Mechanical & Industrial Engineering, and is one of three laboratories that comprise the Human Factors Research Group. CEL was founded in 1992 and is primarily concerned with conducting basic and applied research on how to introduce information technology into complex work environments.

Current CEL Research Topics

CEL has been funded by Atomic Energy Control Board of Canada, AECL Research, Alias|Wavefront, Asea Brown Boveri Corporate Research - Heidelberg, Canadian Foundation for Innovation, Defence Research & Development Canada (formerly Defense and Civil Institute for Environmental Medicine), Honeywell Technology Center, IBM, Japan Atomic Energy Research Institute, Microsoft Corporation, Natural Sciences and Engineering Research Council of Canada, Nortel Networks, Nova Chemicals, Westinghouse Science & Technology Center, and Wright-Patterson Air Force Base. CEL also has collaborations and close contacts with the Mitsubishi Heavy Industries and Toshiba Nuclear Energy Laboratory. Recent CEL projects include:

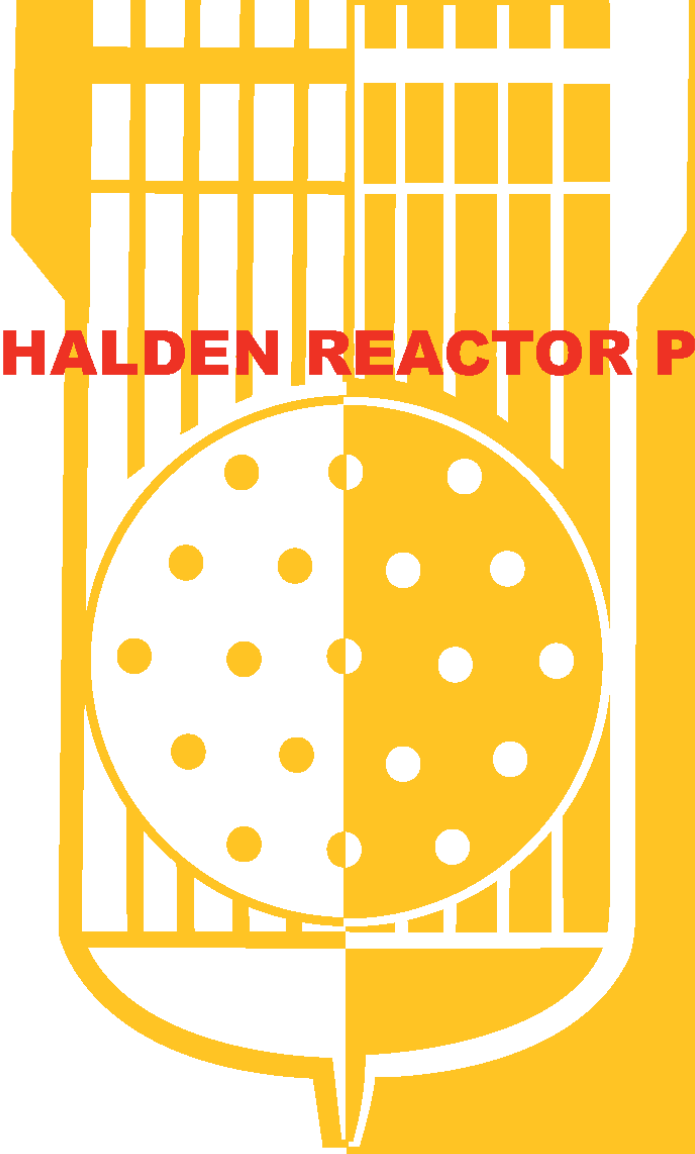
- Developing advanced human-computer interfaces for the petrochemical and nuclear industries to enhance plant safety and productivity.
- Understanding control strategy differences between people of various levels of expertise within the context of process control systems.
- Developing safer and more efficient interfaces for computer-based medical devices.
- Creating novel measures of human performance and adaptation that can be used in experimentation with interactive, real-time, dynamic systems.
- Investigating human-machine system coordination from a dynamical systems perspective.

CEL Technical Reports

For more information about CEL, CEL technical reports, or graduate school at the University of Toronto, please contact Dr. Kim J. Vicente or Dr. Greg A. Jamieson at the address printed on the front of this technical report.

HWR-971

OECD HALDEN REACTOR PROJECT



OECD

The Process Overview Measure: Methodological Developments to Enhance Inter-Rater Reliability

INSTITUTT FOR ENERGITEKNIKK
OECD HALDEN REACTOR PROJECT
P.O. BOX 173, NO-1751 HALDEN, NORWAY
www.ife.no/hrp

HWR-971

OECD HALDEN REACTOR PROJECT

The Process Overview Measure: Methodological Developments to Enhance Inter-Rater Reliability

by

Nathan Lau, Gyrd Skraaning Jr, Tommy Karlsson, Christer Nihlwing, OECD Halden Reactor Project; Greg A. Jamieson, University of Toronto

2011-02-09

FOREWORD

The experimental operation of the Halden Boiling Water Reactor and associated research programmes are sponsored through an international agreement by:

- the Institutt for energiteknikk (IFE), Norway,
- the Belgian Nuclear Research Centre SCK•CEN, acting also on behalf of other public or private organisations in Belgium,
- the Risø DTU National Laboratory for Sustainable Energy, Technical University of Denmark,
- the Finnish Ministry of Employment and the Economy (TYÖ),
- the Electricité de France (EDF),
- the Gesellschaft für Anlagen- und Reaktorsicherheit (GRS) mbH, representing a German group of companies working in agreement with the German Federal Ministry of Economics and Technology,
- the Japan Nuclear Energy Safety Organization (JNES),
- the Korean Atomic Energy Research Institute (KAERI), acting also on behalf of other public or private organisations in Korea,
- the Spanish Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), representing a group of national and industry organisations in Spain,
- the Swedish Radiation Safety Authority (SSM), representing public and privat nuclear organisations in Sweden,
- the Swiss Federal Nuclear Safety Inspectorate ENSI, representing also the Swiss nuclear utilities (Swissnuclear) and the Paul Scherrer Institute,
- the National Nuclear Laboratory (NNL), representing a group of nuclear licensing and industry organisations in the United Kingdom, and
- the United States Nuclear Regulatory Commission (USNRC),

and as associated parties:

- Japan Atomic Energy Agency (JAEA),
- the Central Research Institute of Electric Power Industry (CRIEPI), representing a group of nuclear research and industry organisations in Japan
- the Mitsubishi Nuclear Fuel Co., Ltd. (MNF)
- the Czech Nuclear Research Institute (NRI),
- the French Institut de Radioprotection et de Sûreté Nucléaire (IRSN),
- the Ulba Metallurgical Plant JSC in Kazakhstan,
- the Hungarian Academy of Sciences, KFKI Atomic Energy Research Institute,
- the JSC "TVEL" and NRC "Kurchatov Institute", Russia,
- All-Russian Research Institute for Nuclear Power Plants Operation (VNIIAES), Russia,
- the Slovakian VUJE - Nuclear Power Plant Research Institute, and
- EU JRC Institute for Transuranium Elements, Karlsruhe,

and associated parties from USA:

- the Westinghouse Electric Power Company, LLC (WEC),
- the Electric Power Research Institute (EPRI),
- the Global Nuclear Fuel (GNF) – Americas, LLC and GE-Hitachi Nuclear Energy, LLC, and
- the US Department of Energy (DOE)

The right to utilise information originating from the research work of the Halden Project is limited to persons and undertakings specifically given this right by one of these Project member organisations.

Recipients are invited to use information contained in this report to the discretion normally applied to research and development programmes. Recipients are urged to contact the Project for further and more recent information on programme items of special interest.

Institutt for energiteknikk
OECD HALDEN REACTOR PROJECT

Title

The Process Overview Measure: Methodological Developments to Enhance Inter-Rater Reliability

Author:

Nathan Lau, Gyrd Skraaning Jr, Tommy Karlsson, Christer Nihlwing, OECD Halden Reactor Project; Greg A. Jamieson, University of Toronto

Document ID:
HWR-971

Keywords:
Situation Awareness, Process Overview, Measure, Human Factors, Inter-rater Reliability

Abstract:

The Process Overview Measure assesses monitoring performance by eliciting knowledge about parameter behaviours using the freeze-and-query techniques. Participants from an earlier empirical study reported ambiguities associated with time references in the queries and completeness of the response alternatives, hindering the inter-rater reliability of the measure. An empirical study was conducted to evaluate two methods of improving the inter-rater reliability of the Process Overview Measure. The first method aimed to reduce ambiguities in the time reference by specifying scenarios events to define the time windows for judging parameter behaviours. The second method aimed to improve the completeness of the response alternatives by adding "fluctuating around the same point" as a category of parameter behaviour in the response set. The results indicate that specifying scenario events to define time windows in the queries could improve inter-rater reliability. However, adding "fluctuations" in the response set hindered inter-rater reliability. The results support the use of scenario events to define time windows for judging parameter behaviours in the queries. A revised procedure to the Process Overview Measure is presented.

Issue Date:		Name	Signature	Date
2011-02-09	Prepared by:	Nathan Lau	Sign.	2010-11-30
Confidential grade:	Reviewed by:	Greg A. Jamieson	Sign.	2010-11-30
HRP Only	Approved by:	A. Bye	Sign.	2011-02-09

MAIL	TELEPHONE	TELEFAX		TELEFAX	
P.O. BOX 173	+47 69 21 22 00	Administration	+ 47 69 21 22 01	IND/OC div.	+ 47 69 21 24 90
NO-1751 HALDEN		Nuclear Safety	+ 47 69 21 22 01	VISIT/RID/COSS div.	+ 47 69 21 24 60
Norway		Purchasing Office	+ 47 69 21 24 40	Reactor Plant	+ 47 69 21 24 70

TABLE OF CONTENTS

1. INTRODUCTION

As the operation of nuclear power plants (NPPs) becomes increasing knowledge-based (e.g., see [1]), the notion of Situation Awareness (SA), along with situation assessment, becomes increasingly useful for describing and discussing operator work [2]. However, the application of SA in nuclear process control faces two fundamental challenges. First, SA is poorly defined or, at least, suffers from the lack of a unanimous definition or model as the ambitious scope of the notion aims to cover a vast variety of cognitive work across multiple domains. Second, compounding the lack of consensus about the notion, there is little SA research specific to nuclear process control (or process control in general), especially by comparison to other domains such as aviation [3]. Consequently, the nuclear domain cannot fully capitalize on the SA notion. The OECD Halden Reactor Project (HRP) is addressing the challenges in the application of SA by developing domain-specific characterizations and measures. [4] conceptualizes SA in nuclear process control as consisting of three components – Process Overview, Scenario Understanding, and Metacognitive Accuracy. These components reflect awareness/knowledge acquired through monitoring, diagnosis and self-assessment activities, respectively. [4] further presents measures corresponding to the three SA components. HRP continues to advance SA research in nuclear process control with further development of Process Overview. [5] develops Process Overview by extending the descriptions of the interactions between domain characteristics and monitoring behaviours, and assesses the Process Overview Measure by conducting three empirical studies.

Improving the Process Overview Measure could lead to more accurate and reliable indicator of monitoring performance that would yield better evaluation of control room designs to support operators. This report continues the development of the Process Overview Measure. Specifically, it documents an empirical study evaluating two methodological modifications to the Process Overview Measure as guided by the empirical findings in [5]. The report begins with a review of the Process Overview Measure followed by the experimental method, analysis and results, and discussion of the empirical findings.

2. THE PROCESS OVERVIEW MEASURE

The Process Overview Measure is a domain-specific SA measure that aims to assess operator knowledge/awareness acquired through monitoring. It is developed in accordance with Process Overview – a domain-specific characterization of SA and adapts the Situation Awareness Control Room Inventory (SACRI) [6, 7] and Situation Awareness Global Assessment Technique (SAGAT) [8, 9]. This chapter describes the Process Overview Measure, reviews the current research on the measure, and identifies corollary issues that ultimately lead to the empirical study presented in the subsequent chapters. (Refer to [5] for the full formulation of the Process Overview concept and measure.)

2.1 Description

The Process Overview Measure operationalizes Process Overview as the accurate detection of meaningful changes in relevant process parameters. Process parameters are *relevant* when they effectively represent the operating contexts (e.g., shutdown) and reveal potential process anomalies. Parameter changes are *meaningful* when they represent the systematic trends as opposed to (uninformative) fluctuations.

The Process Overview Measure involves three phases: preparation, data collection and scoring. During the preparation phase of full-scope simulator experiment or evaluation session, process

experts are instructed to perform three inter-connected tasks – development/review of scenarios, selection of relevant parameters, and identification of simulator-freeze points for query administration. These tasks are described in the following paragraphs.

Process experts are typically responsible for developing test scenarios with the characteristics that are useful for the purpose of the study (see [10] for a discussion.) For instance, to evaluate an alarm system for monitoring, the test scenarios must contain process events or faults leading to alarms. The Process Overview Measure does not prescribe any guidance for developing scenarios because the dominant consideration should be the purpose of the empirical study. However, the Process Overview Measure does rely on process experts to develop representative test cases that are relevant to experimental topics and sufficiently challenging to operators.

After the development of the scenarios, process experts select process parameters according to the scenario characteristics. Relying on their knowledge in developing the scenarios, process experts should select a set of *relevant* process parameters that represents the operating context and process events (including faults) in the scenario. In other words, the operators/participants successfully completing the scenarios are expected to know the behaviours of these parameters while monitoring the process. The awareness of these parameter behaviours is elicited through administrations of queries in the form as specified in Figure 1. For empirical studies where process experts do not develop test scenarios but still implement the Process Overview Measure, they must review the scenarios carefully.

Process Overview Query Structure:
Recently, the parameter [code] has:

Process Overview Response Alternatives:
Increased/Stayed the same/Decreased

Figure 1: Query and response format of the Process Overview Measure.

Relevant parameters can typically be classified as (i) context-sensitive or (ii) fault-sensitive according to Process Overview [5]. Context-sensitive parameters reflect the overall plant states based on the operating contexts given to the operators at the beginning of the scenarios. For instance, during start-up at a certain power level, operators often sample a set of key parameters periodically to determine the general progress. The cuing effects for context-sensitive queries should be negligible as these parameters are emphasized during their professional training and work practice. Fault-sensitive parameters reveal the process faults introduced by the scenarios. Therefore, fault-sensitive parameters require close observation. Hence, operators may not sample these fault-dependent parameters during normal operations. Thus, fault-sensitive queries may be subject to cuing effects, prompting consideration of the method by which these queries are introduced (see below). Note that the two classes of parameters could overlap.

Process experts also select timing of simulator freezes in the scenarios to administer queries about the relevant parameters. The timing of these freezes (i.e., administration of the queries) should be determined according to selection of process parameters and scenario characteristics. Some context-sensitive parameters become relevant or irrelevant as the scenarios progress. Context-sensitive queries that are relevant for the entire scenarios may be administered at random times. Fault-sensitive queries often require strategic timing of freezes because the queries need to coincide with the

introduction of the faults without resulting in cuing effects[1]. Two general methods are available to counteract cuing effects of administering fault-sensitive queries. The first method relies on the strategic timing of alarms that occur immediately after the freeze to nullify cuing across participants. The second method relies on administering the queries at the end of the scenarios when the cues from the queries cannot influence operator performance. Either or both method may be employed in any implementation of the Process Overview Measure.

The three tasks performed by process experts in preparation for an experiment - scenario development/review, parameter selection, and freeze identification – are inter-connected and not necessary sequential. For instance, the scenarios may be re-designed to provide effective strategic timing of freezes that can nullify cuing effects. Flexibility across those three tasks should be leveraged to optimise the quality of the SA measurements with respect to the purpose of the empirical studies.

During data collection (i.e., while operators/participants are running the test scenarios), the simulator should freeze according to the timings specified by the process experts during data preparation phase. During simulator freezes, the participants answer the corresponding set of queries without any access to process displays. From here onwards, the participant answers are labelled as “responses”. At the same time, the process experts supporting the data collection answer the queries with access to all the process displays. From here onwards, the process expert answers are labelled as “reference keys”. In addition to collecting the responses and reference keys to the queries, the simulator should log the parameters throughout the scenario for potential verification needs after the experiment.

After the data collection, final scores are the proportion correct (or matches) between the responses and reference keys (collected from the participants and process experts, respectively). These scores may then be analyzed statistically.

2.2 Current research and corollary issues

2.2.1 Empirical Findings on the Process Overview Measure

Current research[2] indicates that the Process Overview Measure is *sensitive* to experimental manipulations [5]. In a Haden Man-Machine LABoratory (HAMMLAB) experiment, Process Overview Measure illustrates the relative advantage between three display types for monitoring NPPs under different operating situations. In another full-scope simulator experiment intended to explore futuristic operational concepts [5], the Process Overview Measure illustrated the impact of transparent automation displays with respect to different staffing configurations on operator monitoring performance. Both experiments also produced other empirical results that corroborated with the Process Overview findings. In summary, the Process Overview Measure revealed the impact of new operational concepts or technology on monitoring performance.

Besides sensitivity, the Process Overview Measure demonstrated acceptable inter-rater *reliability* between process experts. In an empirical study comparing the references keys between three process experts [5], the Process Overview Measure exhibited "fair" to "substantial" agreement (i.e., Kappa between 0.4-0.6), which is reasonable for judgment tasks involving complex information processing (e.g., medical diagnosis as suggested by [11]) as in monitoring NPPs. Note that slightly higher agreement (i.e., Kappa between 0.7-0.8) would be desirable. Two sources of ambiguity are discovered

[1] Cueing effect refers to a shift in participant attention (consciously or unconsciously) due to the appearances or presence of some environmental stimuli. Typically, this effect is unintended in the experiments.

[2] For the purpose of this report, the literature review only focuses on research specific to the Process Overview Measure. General discussion on SA measurement is widely available in the literature (e.g., [26-28]). The literature also contains some discussion on SACRI [6, 7] and SAGAT [8, 9] that inspired the development of the Process Overview Measure.

from discussion with operators and process experts (who provided the responses and reference keys in the empirical studies) during the debriefing sessions. First, the term "Recently" in the Process Overview queries is deemed ambiguous as operators can think of multiple events during the scenario that can be considered important for defining the relevant time period to judge parameter changes. Second, some parameters fluctuate up and down during the scenarios, which are not interpreted as strictly increasing, staying the same or decreasing. These comments prompt a search for methodological improvements in the Process Overview Measure.

2.2.2 Methodological Development

The qualitative and quantitative results from the empirical studies [5] direct research focus towards specifying (i) a clear time reference in the queries, and (ii) comprehensive categories of parameter behaviours in the response set.

To provide a clear time reference in the queries, a potential solution is to specify the starting point of a critical scenario event as the starting time for determining parameter changes as opposed to leaving the judgment (of time) entirely to the operators. Specifying scenario events is more consistent than a time (i.e., number of minutes) with Process Overview [4, 5], which indicates that process operators think in action time, than specifying absolute/clock time in the query [12].

To capture all relevant categories of parameter behaviours, the response set used in [4, 13, 14] requires augmenting with a “fluctuation” option. While the current response set may be conceptually complete from the perspective that any parameter changes must fit into one of the three options (i.e., decreased, stayed the same, and increased), process operators may think in more categories of parameter behaviours that offer additional operational utilities in practice. That is, a parameter may fluctuate abnormally around a value signalling process anomalies. In such cases, even if all operators and process experts know that a parameter is fluctuating, they must select another option given the current Process Overview response set. Some may consider the parameter unchanged while others may judge the parameter to have increased or decreased depending on the exact parameter value at the time of the simulator freeze.

2.3 Objective of the this study

To assess the merits of specifying scenario events for time references and adding fluctuations to the response set, we conducted a controlled experiment to compare the inter-rater agreements between four different variants of the Process Overview Measure.

3. METHOD

3.1 Experimental Design

Each participant performed 56 trials, comprised of fourteen different cases assigned to four treatment combinations (14x4). In total, the data set contained 168 trials from three raters (56x3). The three raters also yielded three combinations of rater-pairs (i.e., Rater A & B, A & C, and B & C) for calculating agreement between any two experts (also see section 4.1).

A 3x2x2 split-plot factorial (SPF) design was employed with a between-subject factor of rater-pairs (AB, AC, and BC), and within-subject factors of time window (Recently/Undefined and Event-based/Predefined) and response alternative (3AFC and 4AFC).

The treatments were completely crossed and counterbalanced by randomizing presentation order of cases and treatments with two restrictions. The first restriction was that each treatment combination occurred only once for each case. The second restriction was that the fourteen different cases were randomly selected without replacement until all cases were selected.

3.2 Participants

The participants were three male process experts. Besides accessibility, one principal selection criteria was their familiarity with two previous empirical studies that investigated the Process Overview Measure. As mentioned in *Procedure and experimental task* (section 3.4), one experiment employing the HAMBO full-scope simulator provided the source data generating the parameter trends. The scenario trials in which participants operated the HAMBO simulator provided the corresponding operating contexts of those parameter behaviours in this study. The second empirical study investigated the inter-rater reliability of the Process Overview Measure that also relied on the full-scope simulator study for its source data. All three process experts participated in the two earlier studies (i.e., Study 2 and 3 in [5]), providing them with the exposure to the simulator, the Process Overview Measure, and scenarios in full-scope simulator experiment. The exposure to the two related studies should prepare these process experts (better than others) to mentally simulate the parameter behaviours with respect to the operating contexts (i.e., scenario descriptions and operator actions given on papers) when they could not observe the development of the nuclear process in real time (i.e., during the full scope experiment). However, the selective recruitment process would mean that "rater" is a fixed as opposed to a random factor (i.e., participant is typically a random factor in both full-scope simulator and inter-reliability studies), limiting the generalization of the results on this particular factor[3].

The relevant characteristics of each process expert are summarized below:

Rater A was a process expert employed at the HRP (for two years). He worked as a control room operator in multiple nuclear plants for fifteen years and participated as a process expert to develop of the HAMBO simulator in late 1990's. For the experiment where the source data for the graphs was drawn in this study, he designed all of the scenarios and advised other HRP scientists on the processes and operations of the physical and simulator plant. He also identified all of the process parameters for the Process Overview queries. During data collection, Rater A played the role of field operator and electrician as required by the scenarios and participant interventions. He was also the designated process expert for completing questionnaires related to participant performance.

Rater B is a recently retired (for around one year) shift supervisor at the nuclear that the HAMBO simulator replicates. He worked as a control room shift supervisor for 32 years and participated in about ten HRP studies prior to this study. He was a participant for the pilot trials of the experiment of the source data, providing additional exposure to the experiment. Furthermore, Rater B acted as the second expert rater (in addition to Rater A) for investigating inter-rater reliability of the Process Overview Measure during the data collection of the full-scope simulator experiment.

Rater C was a process expert employed at the HRP for eleven years. He worked as a control room operator in multiple nuclear plants for fifteen years and participated as a process expert in the development of the HAMBO simulator. Prior to this study, he had similar responsibility for numerous HAMMLAB simulator studies. For the experiment where the source data for the graphs was drawn in

[3] There is no practical solution to the trade off between selective recruitment and generalizablity. Randomly recruited process experts would hinder the validity of the test due to lack of knowledge on the experiment to judge parameter changes. On the other hand, selectively recruited process experts impose some limits on generalization of the statistical results.

this study, he developed the scripts that automatically start-up the simulator plant, implemented the scenarios, co-designed the transparent automation interface and advised other HRP scientists on the processes and operations of the physical and simulator plant. During data collection, he was the designated technical support person for simulator operations. For the earlier inter-rater reliability study, Rater C answered the Process Overview queries after the full-scope simulator experiment (in a set up similar to the treatment combination of Recently/Undefined and 3AFC treatment combination of this study). He responded to queries based on descriptions of scenarios and operator actions as well as parameter trends.

3.3 Experimental manipulations

This experiment included two experimental manipulations – time windows and response alternatives of the Process Overview Measure queries.

3.3.1 Time windows of Process Overview Measure queries

Time window refers to how Process Overview Measure queries frame the time periods for the operators to judge the behaviours of process parameters. This experiment included two types of time windows – (i) Recently/Undefined, and (ii) Event-based/Predefined.

The Recently/Undefined time window (a method employed prior to this study [5]) relied on operators to decide on the time period by using the term “Recently” for judging parameter behaviours (see Figure 1). The operators were instructed to define the start time of the period at the last meaningful plant state and the end time at the simulator freeze for the particular parameter/query.

The Event-based/Predefined time window relied on process experts to predefine the time period by specifying events in the scenarios (during scenario development) for judging parameter behaviours. The query defined the start time at the occurrence of a specific event in the scenario and end time at the simulator freeze to be the period for judging parameter behaviours (see Figure 2).

Since event X (e.g., telephone call from field operators) until now, parameter [code] (e.g., condenser pressure) has:

Figure 2: Process Overview query structure with the Event-based/Predefined time window.

3.3.2 Response alternatives of Process Overview queries

The response alternative refers to the categories of parameter behaviours presented to the operators for answering the Process Overview queries. This experiment included two sets of response alternatives – (i) 3 alternative-forced choice (3AFC) and (ii) 4 alternative-forced choice (4AFC).

The 3AFC response set (a method employed prior to this study) included three categories of parameter behaviours: (i) increased, (ii) stayed the same, and (iii) decreased. The 4AFC response set included four categories: (i) increased, (ii) stayed the same, (iii) fluctuated around the same point, and (iv) decreased.

3.4 Procedure and experimental task

The data collection involved three stages: (i) a 0.5-hour session of introduction and instruction, (ii) eight 1.25-hour sessions of experimental task, and (iii) a 0.5-hour debriefing session.

3.4.1 Introduction and instruction

During the introduction and instruction session, the experimenter verbally described the intent, environment and schedule of the data collection for the study. The instruction of the experimental tasks was delivered in both verbal and written form (Appendix A). The participants were instructed to answer Process Overview queries by inspecting trend graphs generated from logs of a prior full-scope simulator study [5, 13, 14] and considering contexts in which the parameter trends were being developed.

The participants were informed that fourteen cases (i.e., the contexts) were drawn from the prior full-scope simulator study (in which they had participated as process experts). The full-scope simulator experiment collected data from nine NPP crews operating under eight different scenarios (see [13, 14] for descriptions). Eight and six scenarios performed by the second last and the last crews, respectively, made up the fourteen cases in this study. These fourteen cases were selected because comparisons could be made with the dataset collected from a prior reliability study [5] on the Process Overview Measure, which relied on the last two crews of the full-scope simulator study.

Because participants were not situated in the context as the parameters changed (i.e., real time during the full-scope simulator experiment), the instruction explicitly prompted them to consider contextual information as well as scales of the axes when answering Process Overview queries. The Process Overview queries were selected by process experts based on scenario analyses to reflect the expected knowledge from monitoring the nuclear process in those specific scenarios. The explicit reminder was aimed to encourage participants to rely on their expertise in the nuclear process and knowledge of the full-scope experiment that would direct monitoring behaviours during the actual scenario trials. The contextual information was provided on paper through scenario descriptions, task performance scores and brief descriptions of the operator actions with respect to the fourteen cases. The experimenter also informed the participant that they could request the trend graphs to be generated on different scales that deemed appropriate for the parameters in the corresponding contexts. When they were uncertain about the exact sensor of the parameter code, the participant could request access to the simulator in “freeze” mode to identify the sensors pertaining to the specific parameter codes.

After reviewing the instruction, the participants were asked to answer four practice queries corresponding to the four different experimental conditions as described in the section on Experimental task below. Before the data collection sessions began, the participants were asked whether any clarification was necessary.

3.4.2 Experimental task

Data from each participant were collected over eight sessions. Each session took approximately one and a quarter hour, separated by breaks of at least fifteen minutes to avoid fatigue. Each session contained seven cases/trials, each of which contained fourteen to eighteen Process Overview queries[4].

[4]The number of Process Overview queries varies across cases/trials due to characteristics of the scenarios for those cases.

For the treatment of the time window with the level of Recently/Undefined, the participants were required to (i) read the query, (ii) review the graph, (iii) explicitly define the time window by shading in a portion of the graph (with a mouse extending a rectangle; Figure 3), and (iv) select one of the response alternatives with a mouse).

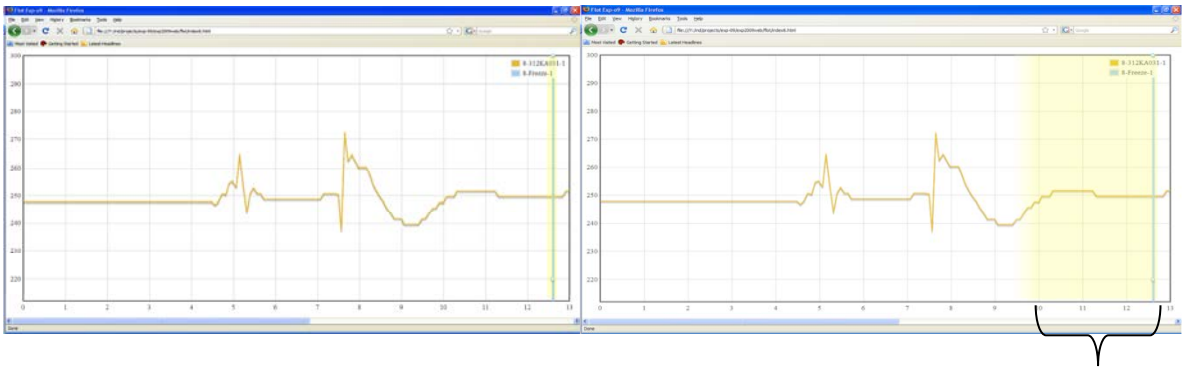


Figure 3: Defining time windows for the Recently/Undefined time window condition.

For the treatment of the time window with the level of Event-based/Predefined, the participants were required to (i) read the query, (ii) review the graph (Figure 4), and (iii) select one of the response alternatives with a mouse. This treatment level did not require operators to shade in any time period for judgment.

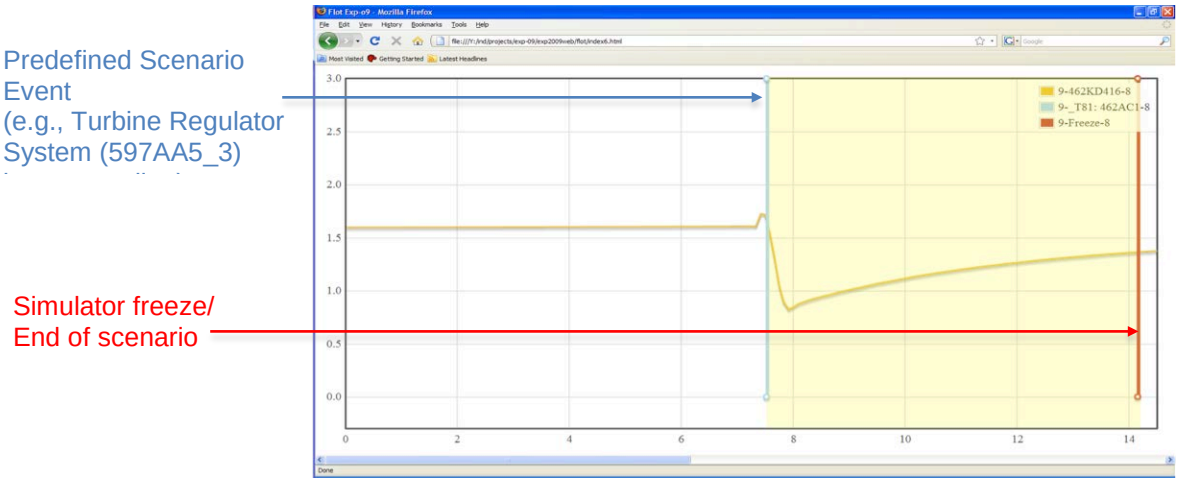


Figure 4: Graph for the Event-based/Predefined time window condition.

For the treatment of response alternative, the 3AFC differed from the 4AFC only in that the response alternative of “fluctuating around the same point” was available for the 4AFC but not the 3 AFC conditions. The two treatments (i.e., experimental manipulations) consisted of two levels each, resulting in four different combinations of experimental manipulations.

Well-defined criteria applicable to all Process Overview queries did not exist for identifying time windows and selecting response alternatives. The participants were explicitly instructed to select a response to a query that would be most meaningful for classifying the parameter behaviour given the specific operating context (i.e., the case). For instance, the response alternative of “fluctuating around the same point” should only be selected if knowledge of the fluctuation offered value to control room operators in operating the NPPs. Small fluctuations with no operational significance should be considered noise and therefore classified as “stay the same”.

3.4.3 Debriefing

After all eight sessions, the data collection ended with a debriefing session to discuss the experimental manipulations in a semi-structured interview format (Appendix B). The debriefing permitted the participants to express their preference and potential improvements to the Process Overview Measure.

3.5 Experimental environment

The data was collected in a large office with two computers, two 30" LCD displays and two 22" LCD displays. One computer was used to present the graphs in PowerPoint files on one connected 30" displays and to collect responses of the participants through the Halden Questionnaire Systems [15] with the other connected to a 30" display. The 22" LCD displays were inactive during the experiments. The second computer was running the HAMBO simulator in "freeze" mode in the background. It was accessible by changing the input signal on one of the 30" LCD displays. The second computer was available in case the participants needed help in recalling the parameter codes with respect to the sensors. (None of participants asked to use the simulator for recalling parameters.) In addition, the participants were provided with all the scenario descriptions and task performance scores with brief descriptions of the operator actions on papers corresponding to the fourteen cases.

3.6 Hypothesis

The two proposed modifications to the Process Overview Measure were expected to improve consistency between raters. First, specifying a scenario event, at which considerations of the parameter behaviours should start, could reduce the ambiguities in comparison to the term "recently". Second, the 4AFC response set could provide a qualitatively distinct option suitable to the behaviours of some parameters thereby increasing agreement between raters over the 3AFC response set (when marginal distribution/chance agreement is factored in).

In addition to hypotheses related to the modifications, the level of agreement between raters could vary based on empirical the findings of an earlier reliability study [5].

4. ANALYSIS AND RESULTS

4.1 Statistical testing

The data analysis involved a two-step statistical procedure (i) calculating Cohen's Kappa, κ (see [16, 17] and Appendix C for a brief review), between raters/participants for each trial[5], and (ii) performing significance testing of differences between κ scores between rater pairs using ANOVA and non-parametric tests. (Note that the κ statistics account for the reduced chance agreement with the increased number of response options.)

The two-step analysis procedure should yield the most accurate results of the collected data. The first step of calculating κ resulted in statistics of agreement between two raters that accounted for the marginal distribution (i.e., chance agreements and number of response options). As mentioned, three

[5] The adopted aggregation procedure, calculating κ between raters per trial, is not ideal as calculating κ between raters per experimental condition most likely produces the best estimates. Furthermore, ANOVA is not the typical significant testing method for κ's (see Appendix C). However, aggregation over experimental conditions (as opposed to trials) limits analysis to graphical inspection and pair-wise comparisons, failing to provide omnibus tests of the full experimental design. Appendix D documents the analysis of κ's based on data aggregation per experimental condition relying on graphical inspection and pair-wise comparisons.

participants led to three rater-pairs and thus three set of κ's (i.e., Rater A & B, A & C, and B & C). The total number of κ's was 168 (3 rater-pairs x 2 levels of time-windows x 2 levels of response alternative x 14 cases). The second step applied ANOVA on κ's to perform significance testing on the full experimental design.

Two Process Overview queries were omitted from data analysis (one in the Recently and 3AFC, and another in the Recently and 4AFC experimental conditions) because of an error with the two graphs. As a result, both 3AFC and 4 AFC in the Recently time window condition contained 225 data points as opposed to 226 data points in the Event-based time window condition.

The κ's were analyzed with ANOVA using Type III/Unique sums of squares with a between-subject factor of rater-pairs (AB, AC, and BC), and within-subject factors of time window (Recently/Undefined and Event-based/Predefined) and response alternative (3AFC and 4AFC).

The multivariate normality assumption was not satisfied. Two of the twelve distributions were not normally distributed. Specifically, the distribution of κ between Rater A and B in the condition of Event-based time window and 4AFC response alternative was negatively skewed; whereas, the distribution of κ between Rater B and C in the condition of Recently time window and 3AFC response alternative was positively skewed. Though generally robust to violations of the normality assumption, ANOVA results would not be particularly robust to distributions heterogeneous in form [18]. Levene's test indicated homogeneity of variance, but there was a *slight* positive correlation between means and standard deviations for the between factors of rater-pairs that could threaten the robustness of the ANOVA results [19]. The sphericity assumption was inherently satisfied as all within-subject treatments only consisted of two levels. From the perspective of the entire data set, the aforementioned characteristics of the raw data were not "gross" violations of ANOVA assumptions. Therefore, the ANOVA results on κ statistics are presented. To be cautious, non-parametric tests [20-22] were employed to verify the subset of the experimental design only containing the significant ANOVA results.

The ANOVA on κ (Table 1) revealed: (a) between-subject main effect of rater-pairs ($F(2,39)=7.85$, $p=0.00$, $\eta^2=.29$, Figure 5); (b) within-subject main effect of time window ($F(1,39)=26.29$, $p=0.00$, $\eta^2=.40$, Figure 6); (c) within-subject main effect of response alternative ($F(1,39)=12.04$, $p=0.00$, $\eta^2=.25$, Figure 8); and (d) interaction effect of rater-pairs and time window ($F(2,39)=3.23$, $p=0.05$, $\eta^2=.14$, Figure 7).

Table 1: ANOVA of Kappa

	SS	Df	MS	F	p	η^2
Fixed, *Between* Effect						
Rater-pair	1.4823	2	0.7412	7.8459	0.00	0.29
Error	3.6841	39	0.0945			
Fixed, *Within* Effects						
Time Window	0.8714	1	0.8714	26.2923	0.00	0.40
Time Window*Rater-pair	0.2140	2	0.1070	3.2280	0.05	0.14
Error	1.2926	39	0.0331			
Response Alternative	0.2714	1	0.2714	13.0350	0.00	0.25
Response Alternative*Rater-pair	0.0853	2	0.0426	2.0479	0.14	0.10
Error	0.8121	39	0.0208			
Time*Window*Response Alternative	0.0101	1	0.0101	0.3821	0.54	0.01
Time*Window*Response Alternative* Rater-pair	0.0301	2	0.0151	0.5690	0.57	0.03
Error	1.0321	39	0.0265			

A series of four non-parametric tests, following the Holm's adjustment procedure of family-wise error rate [20], was performed to verify the significant ANOVA effects. The series of non-parametric tests confirmed all significant ANOVA effects:

- a non-parametric, Kruskal-Wallis One-way ANOVA for the main (between-subject) effect of rater-pair (H(2, N=168)=27.86, p=.00);
- a non-parametric, Wilcoxon Matched Pairs test for the main (within-subject) effect of time window (*Z*=4.93, N=84, p=.00, ES=.54)[6];
- a non-parametric, Kruskal-Wallis One-way ANOVA of rater-pair on the difference scores between Recently and Event-based conditions[7] for the interaction effect of rater-pair and time window (H(2, N=84)=7.74, p=.02); and
- a non-parametric, Wilcoxon Matched Pairs test) for the main (within-subject) effect of response alternative (*Z*=3.08, N=84, p=.00, ES=.34).

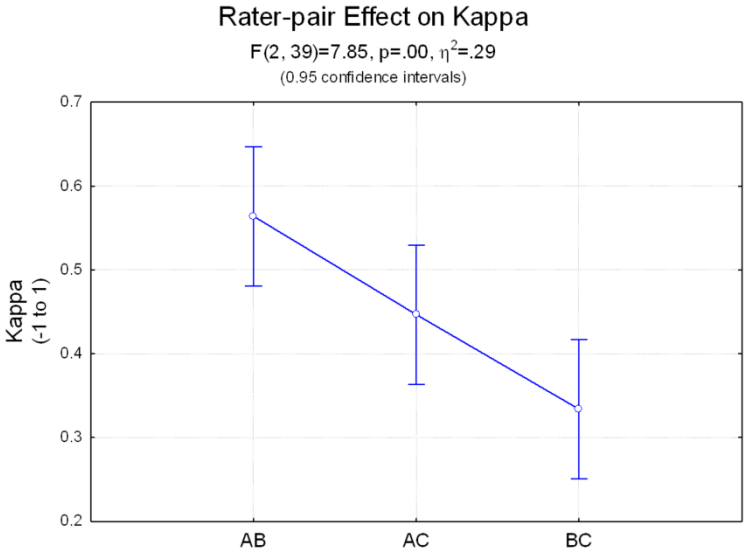
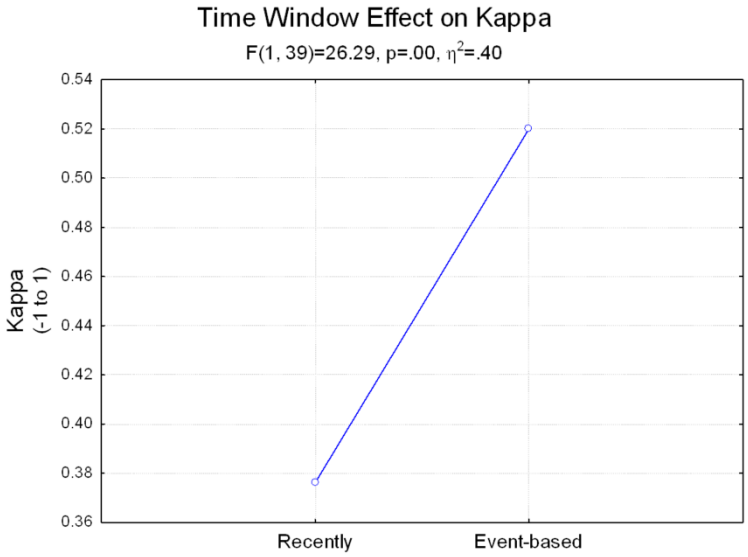


Figure 5: Rater-pair effect on Kappa.

Figure 5 illustrates that raters could bring unique influence in judging parameter changes as discovered in a prior reliability study on the Process Overview Measure [5]. In the prior study, the agreement was highest between raters most involved in the preparation of full-scope simulator study from where the source data was drawn (i.e., Rater A and C). In contrast, the highest agreement in this study was between Rater A and Rater B, who was less exposed to the full-scope simulator experiment than Rater A and C. Figure 7 illustrates the interaction effect between rater-pair and time windows that more precisely describes the influence of raters.

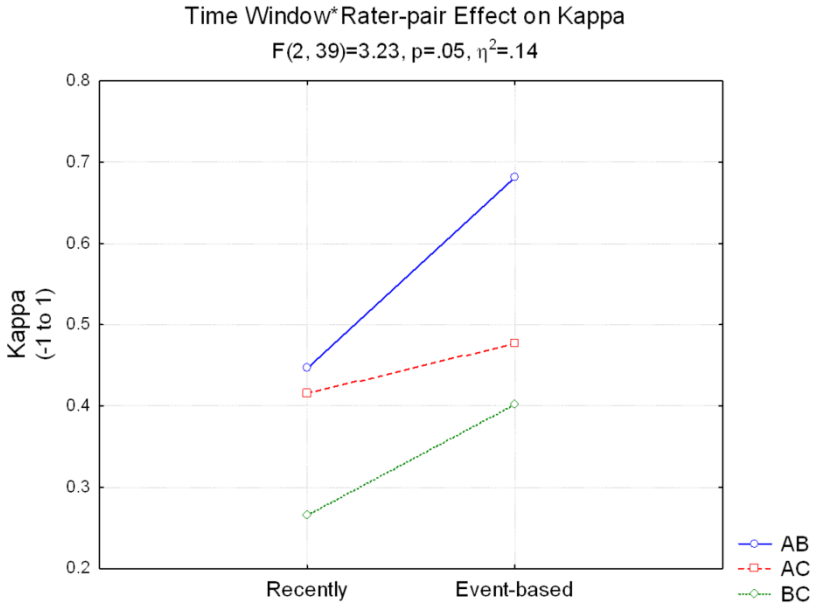
[6] ES denotes effect size for the Wilcoxon Matched Pairs test [22].

[7] As a non-parametric procedure was not available to test interaction effects, the difference scores between Recently and Event-based conditions were calculated, followed by a non-parametric test on rater-pair based on the difference scores. This method of verifying interaction effects is not a standard non-parametric statistical procedure found in the literature. The authors adapt some common statistical techniques to verify the ANOVA effects. This adapted procedure only works for two-way interaction effects in which one of the treatments only has two levels.



Time Windows	Mean	Std Dev.	95% Lower Bound CI	95% Upper Bound CI
Recently/Undefined	0.376	0.027	0.321	0.432
Event-based/ Predefined	0.520	0.028	0.464	0.576

Figure 6: Time window effect on Kappa.

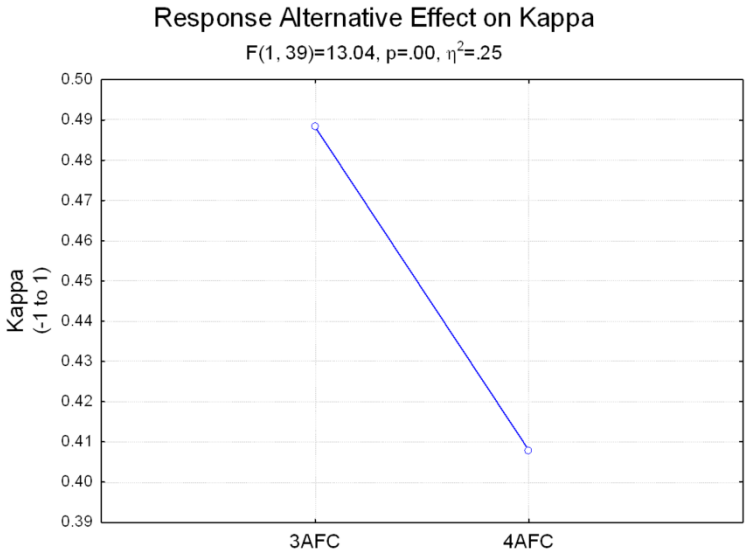


Rater-pairs	Time Windows	Mean	Std Dev.	95% Lower Bound CI	95% Upper Bound CI
AB	Recently/Undefined	0.446	0.048	0.350	0.543
AB	Event-based/ Predefined	0.681	0.048	0.585	0.778
AC	Recently/Undefined	0.416	0.048	0.320	0.512
AC	Event-based/ Predefined	0.477	0.048	0.380	0.574
BC	Recently/Undefined	0.266	0.048	0.169	0.362
BC	Event-based/ Predefined	0.402	0.048	0.305	0.499

Figure 7: Rater-pair and time window effect on Kappa.

Figure 6 illustrates that Recently time windows produced inferior agreement between raters compared to the Event-based ones. Figure 7 further illustrates that all pairs of raters appear to benefit with the

Event-based time windows but to varying degrees. The means of the rater-pairs in the Recently condition resembled the condition of the prior reliability study more than the Event-based condition. Furthermore, the agreement between pairs of raters in the Recently condition also more closely resembled the results of the earlier reliability study than suggested by the main effect of rater-pairs. Specifically, the difference in agreement between Raters A and B compared to Raters A and C was negligible while the agreement between Rater B and C was still the lowest. With respect to the experimental manipulation, Figure 5-Figure 7 illustrates that Event-based time windows could improve agreement between raters as stated in the hypothesis.



Response Alternative	Mean	Std Dev.	95% Lower Bound CI	95% Upper Bound CI
3AFC	0.488	0.026	0.435	0.541
4AFC	0.408	0.026	0.355	0.461

Figure 8: Response alternative effect on Kappa.

Figure 8 illustrates that 3AFC response alternatives produce superior agreement compared to the 4AFC ones in contradiction with the hypothesis.

4.2 Exploratory data analysis, observations and debriefing

The data collected in this study affords additional exploratory data analyses and some conjectural results. The exploratory analyses here involve substantial judgment of the analyst or investigate topics without sufficient experimental control. Although these results must be interpreted with caution, they shed light on the properties of the Process Overview Measure and may guide future research directions.

4.2.1 Judgment in Recently/Undefined time window

For the Recently/Undefined time windows, the participants were required to highlight the area of the trend graphs that defined their windows for judging parameter changes. The analyst coded these data in two ways: (i) the agreement of time windows between raters (i.e., process experts) on a binary scale, and (ii) the starting point of the self-defined time window at 0.5-minute precision. The quantification of these time windows inherently included some subjectivity of the analyst but allowed for a detailed investigation into the Recently/Undefined time window treatment level. Specifically, these

two analyses offer a coarse comparison between inter- and intra-rater consistency in defining Recently/Undefined time windows.

The agreement of the time windows between raters permitted an estimate of inter-rater agreement on defining “Recently”. The proportion agreement on the time windows in the Recently condition between raters (Table 2) was rather low, confirming the findings from the statistical testing above. Note that variation between the starting times as defined by different raters should **not** strictly be interpreted as a continuous scale. For instance, a three-minute difference in start time between two raters may have the same interpretation as a ten-minute difference because both cases suggest that the raters employ different criteria. For this reason, agreement of the time windows between raters as judged by experimenter could be more valid than correlation statistics of time window data quantified at 0.5-minute precision.

Table 2: Proportion agreement on defining time windows in the Recently treatment level.

Rater-pair	Proportion agreement on Recently time-windows
AB	.24
AC	.45
BC	.12

The start times of the windows as defined by the raters permitted another estimate of inter-rater agreement on defining “Recently”. Because time data is on a continuous scale, the intra-class correlations (ICC) statistics [23] are calculated as an estimate of inter-rater agreement on “recently” amongst all three raters. The ICC statistics (ICC(2,1)=.10, ICC(3,1)=.24)[8] are very low, confirming the low agreement between raters in defining “Recently” in the Process Overview queries.

The start times of the windows also permitted estimates of the internal consistency of the raters between the 3AFC and 4AFC response alternative in the Recently/Undefined time window condition. For a continuous scale of time data, ICCs are calculated as estimates of internal consistency[9]. ICC(3,1) for Rater A, B, and C are .51, .41, and .56, respectively. The correlations indicate low to moderate internal consistency within raters in defining “Recently”.

As mentioned, variation in the starting times should not strictly be interpreted as a continuous scale. To study internal consistency of the raters further, the distribution of the differences between the self-defined time windows across the 3AFC and 4AFC conditions are examined under four categories. The first category was the differences of 0 to 0.5 minute, indicating that the rater defined nearly identical time windows for the same two instances of the Process Overview queries. The second category was differences of greater than 0.5 up to 1.5 minutes, indicating that the rater defined practically the same windows with slight differences most likely due to noise. The third category was differences of greater than 1.5 up to 2.5 minutes, indicating that the rater might have defined the same or different time windows (i.e., equivocal or grey areas). The last category was difference of greater than 2.5 minutes, indicating that the raters had defined different time windows. Based on Table 3, the raters appeared to be consistent approximately 70% of the time in defining time windows for the same two instances of the Process Overview queries. Table 3 suggests slightly higher internal consistency within raters than the ICCs on defining “recently”. Internal consistency of approximately 70% agreement may be considered adequate for defining “recently”.

[8] ICC(2,1) treats raters as a random factor; whereas, ICC(3,1) treats raters as a fixed factor. In other words, ICC(2,1) is an estimate of reliability independent of the participants/raters recruited for the experiment; whereas ICC(3,1) is an estimate of reliability the specific raters in this data set.
[9] ICC(2,1) is not presented as the internal consistency estimate is only applicable for the particular judge.

Table 3: Intra-rater difference in starting times of the time windows for the same queries (between 3AFC and 4AFC response alternative in Recently time window conditions).

	Rater A		Rater B		Rater C	
Difference between start time of the self-defined time windows between 3AFC and 4AFC conditions (Δ min)	Freq	Cumulative %	Freq	Cumulative %	Freq	Cumulative %
0-0.5 (no difference)	111	49.33	96	42.67	87	38.67
1-1.5 (slight differences that were likely due to noise)	46	69.78	61	69.78	65	67.56
1.5-2.5 (differences that might be due to noise)	14	76.00	30	83.11	19	76.00
>2.5 (significant differences that were unlikely due to noise)	54		38		54	
Total number of queries	225					

Cursory comparisons between inter-rater agreement and internal consistency estimates suggest that the raters appeared more consistent with themselves than one another in defining time windows in the Recently condition. In other words, the process experts are likely to be more consistent to themselves than to one another in defining the time windows in the Recently condition.

Reviews of the time windows and discussions with raters (i.e., process experts) suggested that defining the start time of the windows for judging parameter behaviours may be based on many considerations such as the following:

- start time at the beginning of the scenario (or scenario period) when scenario events do not directly affect the parameters
- start time at the last important event even though the scenario event does directly affect the parameters
- start time at the beginning (antecedent) of an relevant event because experts judge that knowledge about parameter behaviours from time before the event to the time of freeze is critical
- start time at the end (consequence) of an relevant event because experts judge that knowledge about parameter behaviours from the time after the event to the time of freeze is critical
- start time at “a few minutes” before the freeze because experts define “recently” based on their unique operational experience with other control room operators

In the debriefing discussions, the process experts often emphasized the importance of scenario characteristics in defining “recently”. It appeared that experts relied on their mental models of the scenario more than inspecting the trend graphs in defining “recently”.

During data collection of this study, a question from both Raters A and C, who prepared part of the experiment where the source data was drawn, exemplified their reliance on knowledge about the simulator and scenarios. Specifically, Raters A and C identified that the trend graph for one parameter appeared to be inconsistent with the scenario development. The process experts quickly resolved their own question by realizing that a unique event in the scenario prevented the sensor from accurately measuring the values of that parameter. The concern raised and ultimately resolved by Raters A and C themselves highlighted the level of expertise and judgment involved in assessing parameter behaviours even in judging parameter behaviours from graphs.

4.2.2 Intra-rater consistency

The collected data permitted three imprecise estimations of intra-rater consistency. The data collection environment and procedures were not intentionally designed to provide such estimates, potentially leading to confounding factors in the results. Nevertheless, the intra-rater consistency estimates could provide insights on future research directions.

As presented in the section above, two indicators of intra-rater consistency are derived from the comparison of time windows defined in the treatment level of Recently between the 3AFC with the 4AFC response alternative conditions. First, the intra-class correlations demonstrate moderate level of internal consistency (i.e., 0.4 < ICC(3,1) < .6). Second, based on the distributions of differences between the self-defined time windows across two experimental conditions, raters defined almost identical time windows in the 3AFC as in the 4AFC conditions for 70% of the queries (Table 3). The raters appeared reasonably consistent in defining "recently" across the two response alternative conditions.

The third indicator of rater consistency was the degree of agreement for data collected from the same raters between this study and a prior reliability study [5] in the treatment combination of 3AFC response alternative and Recently time window. In the prior reliability study, Rater A and B answered the same Process Overview queries in real time during data collection of a full-scope simulator study; whereas, Rater C answered the Process Overview queries in the same office of this experiment based on trend graphs after the full-scope simulator experiment. Despite different data collection procedures and environment between the two studies, the raters demonstrated "moderate" agreement (see Appendix C for interpretation) with themselves (Table 4). Furthermore, the differences between κ's of the raters were negligible (Figure 9).

Table 4: Intra-rater agreement between two reliability studies.

Rater	Intra-rater agreement	
	κ	σ_κ
A	.543	.049
B	.572	.048
C	.604	.050

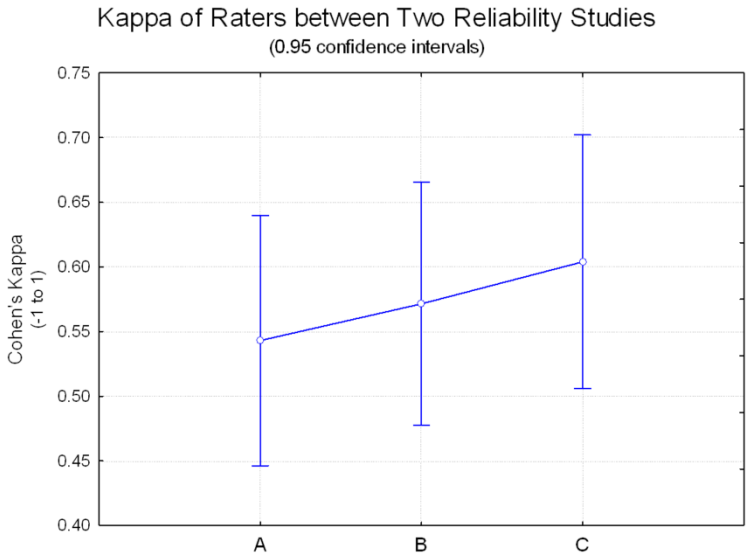


Figure 9: Intra-rater consistency across two reliability studies in Recently time windows and 3FAC response alternative treatment conditions.

In summary, the preliminary evidence indicates reasonably consistent judgment within process experts in answering Process Overview queries.

4.2.3 Preference of process experts

All three process experts indicated that the two proposed modifications to the Process Overview Measure were useful. They find that the Event-based time windows reduced ambiguities. Furthermore, they believed that recalling major events in the scenarios was acceptable when the participants would not have access to the trend graphs to respond to Process Overview queries during data collection. Their opinions on the treatment of time windows converged with the quantitative results.

For the treatment of response alternatives, the process experts believed that the extra option in the 4AFC condition could improve diagnositicity of operator monitoring performance (i.e., Process Overview). One process expert mentioned that description of the new response alternative should be revised slightly. "Fluctuating around the same point" gave the impression of parameters moving above and below a value. However, some fluctuations would be better described as varying above and below some values. Therefore, a better description could be "fluctuating near a point". Although expert opinions and the rationale for including the additional response alternative were consistent, the overall statistical results indicated that the experts implicitly disagreed on what defines fluctuations (see Figure 8).

5. DISCUSSION

5.1 Proposed modifications to the Process Overview Measure

This study intended to evaluate the effectiveness of (i) predefining time windows in the queries with events in the scenarios, and (ii) introducing a new response alternative to include parameter fluctuations to improve inter-rater agreement of the Process Overview Measure.

The results indicate that predefined time windows using scenario events (i.e., Event-based/Predefined time windows) improve agreement between process experts over undefined time windows (i.e., Recently/Undefined). During debriefing, the process experts/participants also approved of the modification. This finding is consistent with the hypothesis. Therefore, Process Overview queries should include a scenario event as the start time for eliciting operator awareness of parameter behaviours during simulator freezes.

Process Overview [4, 5] provides the conceptual basis for proposing the use of scenario events as a method to specify or anchor time periods in queries. Defining time periods according to key events in scenarios is operationally and psychologically consistent with the information processing behaviour of process operators according to field observations (e.g., [12]). Though employing absolute time scales (e.g., minutes) appears to provide an objective solution, operators often psychologically experience time from an operational perspective. In summary, the observed improvement in inter-rater agreement using scenario events to define time windows is justified both conceptually and empirically.

The exploratory analysis further illuminates the main results of this study and Process Overview (the concept). The process experts tend to define "recently" or time windows relatively consistently across two experimental conditions (i.e., 3AFC vs. 4AFC) but disagree with each other considerably, confirming that undefined time windows are indeed one factor challenging inter-rater agreement. The exploratory analysis further indicates that perception of parameter behaviours during monitoring involves substantial information processing and expertise. The exploratory analysis confirms that

process experts do consider many contextual factors in determining parameter behaviours from trend graphs (e.g., scenario descriptions) even situated outside the control room environment. The number of (at least seemingly[10]) legitimate considerations for defining "Recently" exemplifies such complex and implicit processing.

The results indicate that the additional response option of fluctuations (in the 4AFC conditions) hinders agreement between process experts over the original response alternative set (i.e., 3AFC). However, during debriefing, the process experts advocate the new response option as meaningful for monitoring. In brief, the quantitative results contradict the hypothesis and expert opinions.

According to debriefing discussions in this study (and personal communication prior to the study), the process experts agree on the conceptual usefulness of knowing fluctuations for various parameters. However, the empirical results suggest that the process experts appear to disagree, at least implicitly, on the criteria that define fluctuations. The lack of agreement on the defining characteristics of fluctuations may be a result of the inherent fuzziness related to the behavioural category. Fluctuations include increasing and decreasing values that indicate no practical changes on average. In other words, underlying characteristics of fluctuations may rely on all three of the other behavioural categories. The conceptual distinctness of fluctuating parameters in the mind of process experts is difficult to transfer onto a measurement scale in a manner that improves reliability and diagnosticity.

The exploratory analysis suggests that the process experts can answer Process Overview queries fairly consistently across different data collection settings (e.g., real time vs. post experiment). The extensive reliance on knowledge about the simulator/nuclear power plant and scenarios may have minimized the significance of the data collection environment. Though data were not collected with sufficient control for studying the impact of data collection environment in answering Process Overview queries, there are indications that post-experiment scoring with process experts is an acceptable alternative when real-time scoring becomes unfeasible.

5.2 Implication: Modification to the Process Overview Measure

The empirical results indicate that inter-rater agreement of the Process Overview Measure could improve by specifying an important process event in the scenario (or scenario period) as the starting time for judging parameter behaviours of Process Overview queries (i.e., the Event-based/Predefined time window treatment level). Therefore, future implementation of the Process Overview Measure should predefine time windows using scenario events in the queries (Figure 10).

Process Overview query:

Since event X (e.g., telephone call from field operators) until now, parameter [code] (e.g., condenser pressure) has:

Process Overview response alternatives:

(i) decreased (ii) stayed the same (iii) increased

Figure 10: Recommended version of query structure and response set for the Process Overview Measure.

[10] In many cases, it is not feasible to assess the correctness of subject matter experts.

Predefining time windows in queries is an additional step to the procedure of the Process Overview Measure. While developing scenarios for experiments or training sessions, process experts not only need to identify parameters but also construct or select scenario events to develop Process Overview queries. The scenario events should have two qualities. First, the events should be salient and central to the scenarios (or scenario periods) because the detection of the events should not become part of the assessment. In other words, all operators should be aware of the events. Examples include alarms of major systems, scram of the simulated NPP and telephones calls from field operators about equipment failures. Second, the events should be robust or independent to operator control actions because queries would become inapplicable in cases where operators prevent the intended events from occurring during the scenarios. When the intended events predefining the time windows do not occur, the queries and responses could lead to invalid measurements of Process Overview.

In addition the procedural changes to the developing Process Overview queries, process displays in the observation gallery and the data logging systems for the simulator should be updated to store and present the scenario events, respectively. The updated process displays could support process experts in judging parameter behaviours and the logging system could offer the possibility of post-experiment/data collection scoring by process experts.

5.3 Limitation

The method of this study aimed to facilitate the kinds of information processing and judgment similar to those necessary in answering Process Overview queries during data collection of full-scope simulator experiments. The exploratory data analysis demonstrates evidence of success in eliciting such complex information processing in this study. Thus, the effects of the experimental manipulations revealed in this study are expected to generalize to implementations of the Process Overview Measure for full scope simulator studies.

Both the method and data analysis, however, could not ensure that the findings are completely based on process experts or participants being consistently engaged in the same information processing or cognitive activities as in full-scope simulator environment. Relying on trend graphs and paper records to simulate parameter behaviours in dynamic situations for answering Process Overview queries could be challenging even for process experts. Thus, the process experts could be tempted to convert the experimental task, which should involve a mix of visual inspection of parameter trends and mental simulation of the contexts, into a graphical judgment task. The Event-based/Predefined condition appears particularly prone to the temptation of simplifying the experimental task. The exploratory data analysis does not indicate that the process experts have merely performed a graphical judgment task, but modification of the Process Overview Measure requires validation in full-scope simulator experiments.

5.4 Future work

Future work is necessary to validate the recommended modification of the Process Overview Measure in a representative operational environment. Inter-rater reliability studies that employ representative operational settings and multiple process experts involved in experiment preparation could validate the success of the modified Process Overview Measure. Furthermore, the feasibility of identifying events during scenario development that set the time windows for judging parameter behaviours is unknown until the modified Process Overview Measure is tried in a full-scope simulator experiments. Qualitative data from control room operators serving as participants in the full-scope simulator experiments should also be collected for further assessment of the modified Process Overview Measure.

Future work may investigate operator conceptualization of "fluctuation" in details. The empirical evidence suggests that the definition or concept of fluctuation is complex in operational settings. To successfully operationalize the idea of fluctuation, qualitative research appears necessary to fully understand the defining characteristics of fluctuating parameters and the operational utility of knowing such fluctuations.

6. CONCLUSION

The objective of this study was to improve an assessment tool for operator monitoring performance - the Process Overview Measure. The empirical study investigated two proposals: (i) predefining time windows in the query with scenario events, and (ii) adding the "fluctuating around the same point" as a category of parameter behaviours in the response set. The experiment indicates that predefining time-windows with scenarios events could improve agreement between process experts; whereas adding the "fluctuating around the same point" as a response alternative could hinder agreement (relatively to the version of the Process Overview Measure prior to this study). Therefore, the Process Overview Measure now requires process experts to specify events to be the starting times for judging parameter behaviours when they identify relevant process parameters for the queries during scenario development.

The improved inter-rater agreement between process experts for the new Process Overview Measure should reduce noise in Process Overview measurements, thereby providing more sensitive and reliable indicator of operator monitoring performance. Therefore, the improvement of the Process Overview Measure should contribute to the assessment of operator performance and control room support in monitoring NPPs.

7. REFERENCES

[1] K. J. Vicente, *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. Mahwah, N.J.: Lawrence Erlbaum Associates, 1999.

[2] N. Moray, "Oú sont les neiges d'antan?," in *Human Performance, Situation Awareness and Automation*, Daytona Beach, FL, 2004.

[3] J. Patrick and N. James, "A task-oriented perspective of situation awareness," in *A Cognitive Approach to Situation Awareness: Theory and Application*, S. P. Banbury and S. Tremblay, Eds., Hampshire, UK: Ashgate, 2004.

[4] G. Skraaning Jr*., et al.*, "The Ecological Interface Design Experiment (2005)," OECD Halden Reactor Project, Halden, Norway HWR-833, 2007.

[5] N. Lau*, et al.*, "Situation Awareness in Monitoring Nuclear Power Plants: The Process Overview Concept and Measure," OECD Halden Reactor Project, Halden, Norway HWR-954, 2010.

[6] D. N. Hogg*, et al.*, "Development of a situation awareness measure to evaluate advanced alarm systems in nuclear power plant control rooms," *Ergonomics,* vol. 38, pp. 2394-2413, 1995.

[7] D. N. Hogg*, et al.*, "Measurement of the Operator's Situation Awareness for Use Within Process Control Research: Four Methodological Studies," OECD Halden Reactor Project, Halden, Norway HWR-377, 1994.

[8] M. R. Endsley, "Direct Measurement of Situation Awareness: Validity and Use of SAGAT," in *Situation awareness: analysis and measurement*, M. R. Endsley and D. J. Garland, Eds., Mahwah, NJ: Lawrence Erlbaum Associates, 2000, pp. 147-174.

[9] M. R. Endsley, "Measurement of situation awareness in dynamic systems," *Human Factors,* vol. 37, pp. 65-84, 1995.

[10] G. Skraaning Jr., "Experimental Control versus Realism: Methodological Solutions for Simulator Studies in Complex Operating Environments.," OECD Halden Reactor Project, Halden, Norway HPR-361, 2003.

[11] J. Sims and C. C. Wright, "The Kappa Statistics in Reliability Studies: Use, Interpretation, and Sample Size Requirements," *Physical Therapy,* vol. 85, pp. 257-268, 2005.

[12] V. de Keyser, "Structuring of Knowledge of Operators in Continuous Processes: Case Study of a Continuous Casting Plant Start-up," in *New Technology and Human Error*, J. Rasmussen*, et al.*, Eds., Chichester, UK: John Wiley & Sons Ltd., 1987, pp. 247-259.

[13] M. H. R. Eitrheim*, et al.*, "Staffing Strategies in Highly Automated Future Plants Results from the 2009 HAMMLAB Experiment," OECD Halden Reactor Project, Halden, Norway HWR-938, 2010.

[14] G. Skraaning Jr*, et al.*, "Coping with Automation in Future Plants: Results from the 2009 HAMMLAB Experiment," OECD Halden Reactor Project, Halden, Norway HWR-937, 2010.

[15] A. Drøivoldsmo, "New tools and technology for the study of human performance in simulator experiments," PhD dissertation, Norwegian University of Science and Technology, Trondheim, Norway, 2003.

[16] J. Cohen, "Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," *Psychological Bulletin,* vol. 70, p. 4, 1968.

[17] J. Cohen, "The coefficient of agreement for nominal scales," *Educational and Psychological Measurement,* vol. 20, pp. 37-46, 1960.

[18] R. E. Kirk, *Experimental design: procedures for the behavioral sciences*, 3rd ed. Pacific Grove, CA: Brooks/Cole, 1995.

[19] Statsoft. (2010, May). *ANOVA/MANOVA*. Available: http://www.statsoft.com/textbook/anova-manova/#homogeneity

[20] D. Howell, *Statistical Methods for Psychology*, 5th ed. Pacific Grove, CA: Duxbury/Thomson Learning, 2002.

[21] A. Field and G. Hole, *How to design and report experiments*. Thousand Oaks, CA. USA: Sage publications Ltd, 2003.

[22] G. W. Corder and D. I. Foreman, *Nonparametric statistics for non-statisticians: a step-by-step approach*. Hokoken, NJ, USA: John Wiley & Sons, Inc., 2009.

[23] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bulletin,* vol. 86, pp. 420-428, 1979.

[24] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics,* vol. 33, pp. 159-174, 1977.

[25] P. E. Shrout, "Measurement reliability and agreement in psychiatry," *Statistical Methods in Medical Research,* vol. 7, pp. 301-317, 1998.

[26] E. Jeannot, "Situation Awareness, Synthesis of Literature Research," EUROCONTROL, Brussels, Belgium ECC Note No. 16/00, 2000.

[27] E. Jeannot*, et al.*, "The Development of Situation Awareness Measures in ATM Systems," Eurocontrol, Brussels, Belgium HRS/HSP-005-REP-01, 2003.

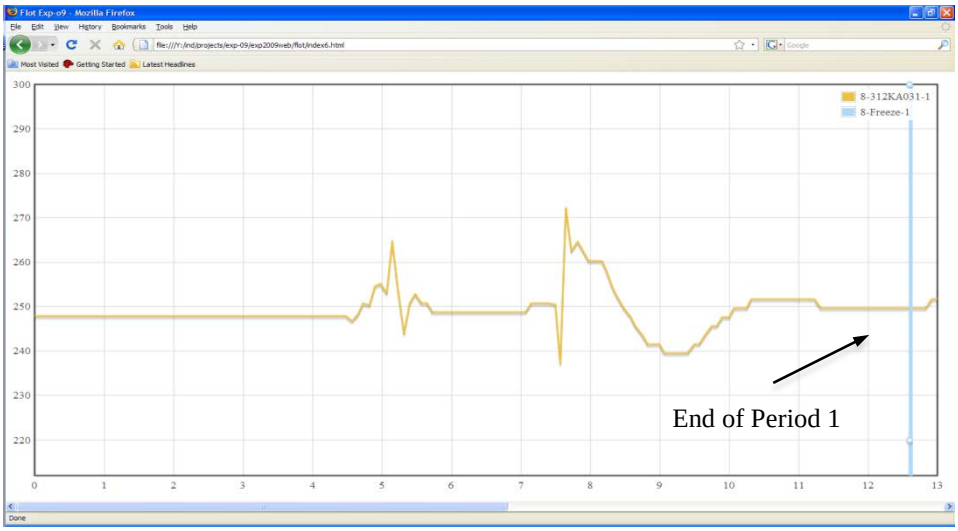
[28] R. Breton and R. Rousseau, "Situation Awareness: A review of the Concept and its Measurement," Defence Research and Development Canada, Valcartier, QC, Canada TR 2001-220, 2001.

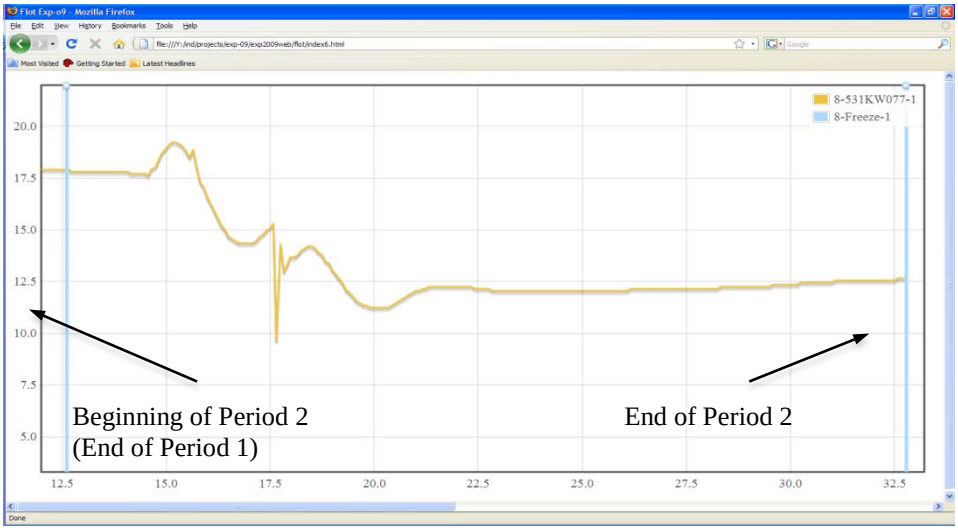
APPENDIX A: INSTRUCTION

Introduction

The aim of this study is to assess inter-rater reliability of different modifications to the Process Overview Measure. The experimental task requires inspecting graphs of various parameters and responding to multiple-choice questions about the changes of those parameters.

A typical graph looks like the following:





You will see one line for the first scenario period and two lines for the second scenario period at the ends of each graph. These lines define the scenario periods.

The following two types of questions are administered:

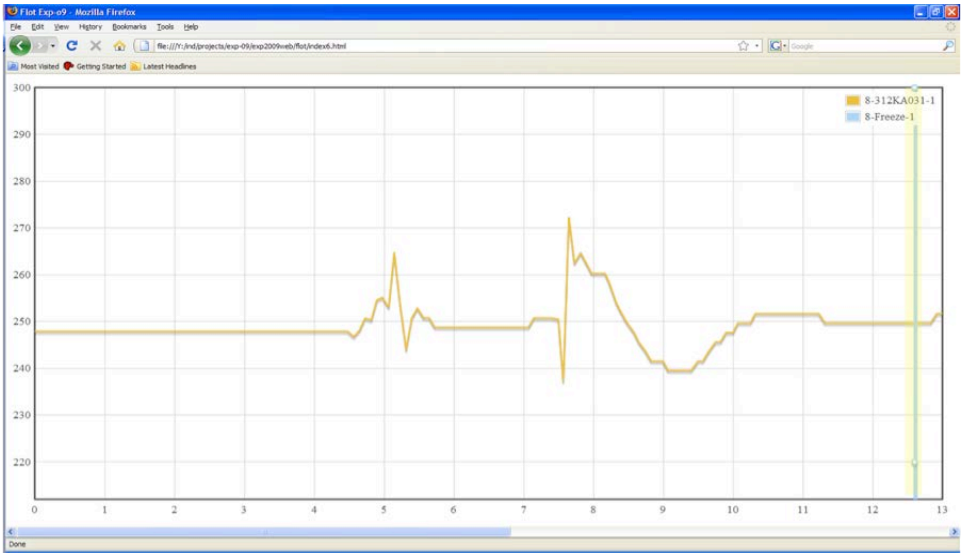
Type 1:

I den senaste tiden har summa drivdonsprocent 221KW700 (i %):

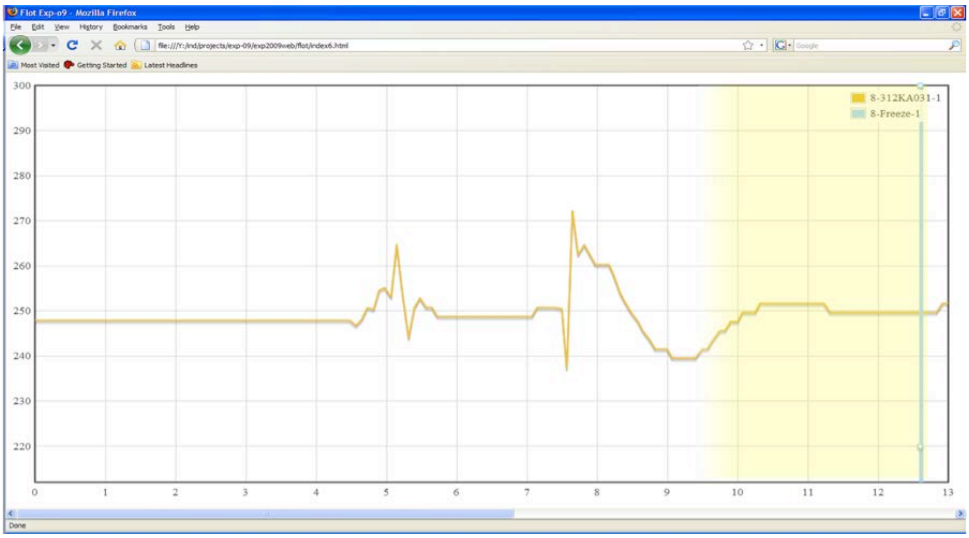
Type 2:

Från sista gången den automatiska agenten stoppade i den här perioden tills nu, har HC-flödet 211KW032:

When the question begins with “I den senaste tiden”, the graph looks like the following:



You should first click on the yellow bar/box and increase the width of the box (by dragging) to the point that covers the portion of the graph that you find **meaningful** to make your judgment about the change for that parameter (like below).

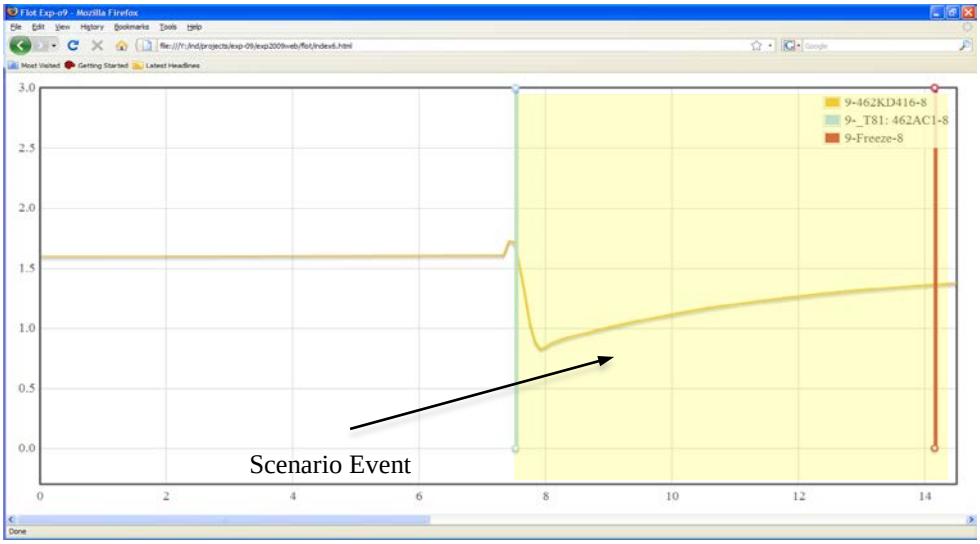


There are many factors in defining a “recent” time period that are meaningful to answer the question. The following is a non-exhaustive list of considerations:

- Scenario characteristics/descriptions
  - Process faults
  - Process events
- Operator actions (or lack thereof) according to OPAS
- Automation actions
- Alarms
- Parameters characteristics

Avoid defining the “recent” period solely by visual inspection of the graphs. It is important that the time period concerning “recently” is operationally meaningful to you.

When the question begins with a scenario event (e.g., “Från sista gången den automatiska agenten stoppade i den här perioden tills nu”), you should see an extra line on the graph that signifies the event specified in the question. For these questions, make the judgment about the parameter change between the time of the event to the end of the scenario period (as depicted by the two lines and highlighted yellow).



There are two sets of multiple-choice responses to the questions (which should be self-explanatory).

(1) minskat (2) varit konstant (3) ökat

(1) minskat (2) varit konstant (3) varierat kring samma punkt (4) ökat

Like defining recently, there are many factors to detecting **meaningful parameter changes**. In making your judgment, the following is a non-exhaustive list of considerations:

- Scenario characteristics/descriptions
  - Process faults
  - Process events
- Operator actions (or lack thereof) according to OPAS
- Automation actions
- Alarms
- Parameters characteristics (e.g., typical noise for the sensor)

Be aware of the scale (y-axis). If necessary, ask for the experimenter to rescale the graph for you.

Avoid judging the changes sole by visual inspection of the graph. It is important that the parameter change is operationally meaningful to know (rather than visually detectable). Note that this applies to the differences between the choice of ”varit knostant” and “varierat kring samma punkt”. If the fluctuation is very small without any operational significance or value, it is more meaningful to respond “varit konstant”. However, if the knowledge about the fluctuation is valuable for the operation, then “varierat kring samma punkt” is more meaningful and appropriate.

APPENDIX B: DEBRIEFING GUIDE

Agenda

1. Thank participants
2. Semi-structured interview
3. Ask participants for any other feedback and questions.
4. Inform participants to contact experimenter/HRP for any further requests/concerns.

Time Windows

Which type of query you rather have: with or without events?

Why?

Extra questions if necessary:

What do you think of the event probe in specifying a time window?

Does it help your definition of the queries or hinder your flexibility in judgment?

Do you have some examples where the event probe help and hinder you?

How do you think of time while operating nuclear power plants?

Response Alternatives

Which type of response options do you rather have: the 3AFC or 4AFC?

Why?

Extra questions if necessary:

What do you think of the additional response option for judging parameter changes?

Do you have cases where the additional or the lack of the option makes it difficult or confusing to answer the query?

Is the response set complete for operating nuclear power plant?

APPENDIX C: DESCRIPTION OF COHEN'S KAPPA

The Cohen's Kappa, κ, [11, 16, 17] is applied to assess the agreement between raters on a nominal scale. Equation 1 and 2 express κ and its standard error, σ_κ, respectively.

$$\kappa = \frac{p_o - p_c}{1 - p_c} = \frac{f_o - f_c}{N - f_c}, \quad [1]$$

$$\sigma_\kappa = \sqrt{\frac{p_o(1-p_o)}{N(1-p_c)^2}} = \sqrt{\frac{f_o(N-f_o)}{N(N-f_o)^2}}, \quad [2]$$

where (for both Equation 1 and 2) p_o=proportion of units agreed; f_o= frequency of units agreed; p_c=proportion of units agreed expected by chance; f_c=frequency of units agreed expected by chance; N denotes total sample size.

κ ranges from -1 to 1, for which the lower bound expresses complete disagreement and upper bound express complete agreement. The statistic is similar to basic correlation statistics except that κ is adjusted for the agreement by chance.

Cohen also provides Equation 3 for significance testing between two κ's :

$$z = \frac{\kappa_1 - \kappa_2}{\sqrt{\sigma_{\kappa_1}^2 + \sigma_{\kappa_2}^2}} \quad \textbf{[3]}$$

Table 5 presents the commonly accepted interpretation of κ [11] (also see e.g., [24, 25]). For tasks in complex environments involving substantial judgment (e.g., medical diagnosis), κ's above 0.4 is acceptable and generally do not exceed 0.7.

Table 5: Commonly accepted interpretation of Cohen's Kappa.

Kappa coefficient	Strength of agreement
$\kappa \leq 0$	Poor
$0 < \kappa \leq .2$	Slight
$.2 < \kappa \leq .4$	Fair
$.4 < \kappa \leq .6$	Moderate
$.6 < \kappa \leq .8$	Substantial
$.8 < \kappa \leq 1$	Almost perfect

APPENDIX D: KAPPA ANALYSIS WITH ALTERNATIVE DATA AGGREGATION PROCEDURE

Table 6 and Figure 11 summarize the results in terms of κ calculated per experimental conditions[11] (see Appendix C for a formula for calculating κ statistics).

Table 6: Cohen's Kappa calculated per experimental conditions.

	AB		AC		BC	
	κ	σ_κ	κ	σ_κ	κ	σ_κ
Recently & 3AFC	0.497	0.050	0.467	0.052	0.308	0.056
Recently & 4AFC	0.404	0.046	0.429	0.045	0.231	0.048
Event & 3AFC	0.780	0.035	0.541	0.048	0.424	0.048
Event & 4AFC	0.610	0.041	0.431	0.045	0.408	0.044

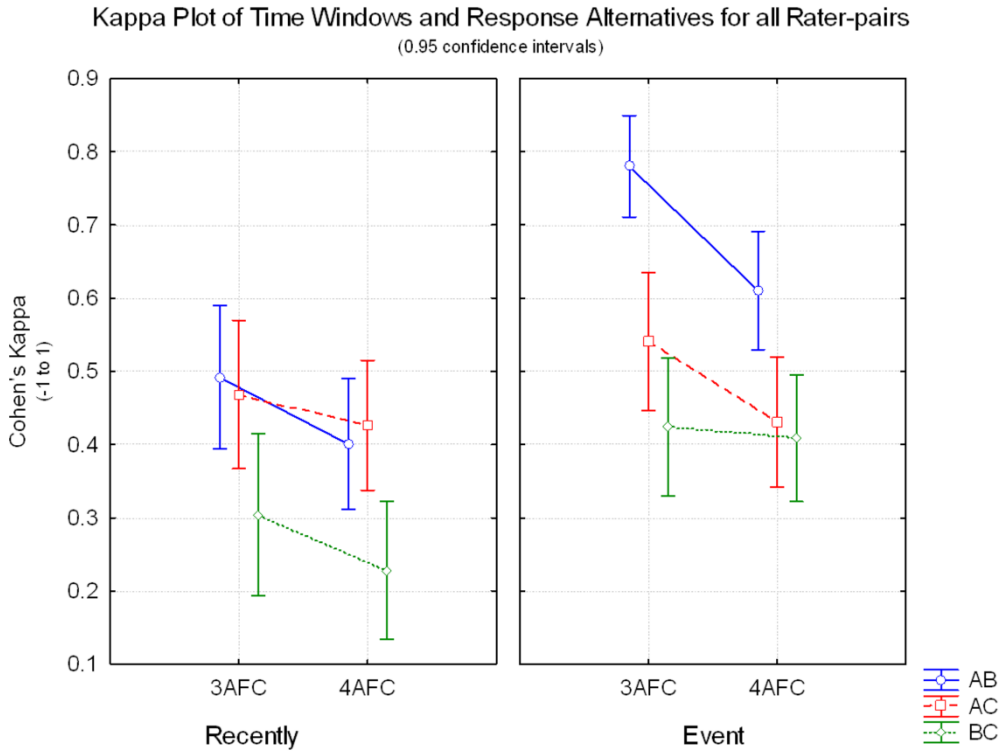


Figure 11: Plot of kappa calculated per experimental conditions.

To varying degrees, the rater-pairs demonstrated similar treatment effects - higher agreement with Event-based time windows and 3AFC response alternatives than Recently time windows and 4AFC response alternatives, respectively. Given the general consistency between raters, κ's were calculated neglecting rater-pairs and interactions to examine main effects only.

[11] Note that the data aggregation and confidence intervals for Cohen's κ have different properties from some other scales (e.g., response time). The data aggregation and variance calculation follow Cohen's formula in Appendix C as opposed to the standard mean and variance formula for ANOVAs. The data aggregation and variance calculation are intended for pair-wise comparisons using the Z-test (see Appendix C). For this reason, the means and confidence intervals are plotted for statistical inferences, even though the data set is treated as repeated-measures/RBF design. Typically, graphical plots of means and confidence intervals for measurements other than κ does not lead to correct statistical inferences for repeated-measures design because the confidence intervals does not appropriately account for within-subject variance.

Following the Holm's family-wise error-rate adjustment for multiple comparisons [20], Cohen's significance testing procedure (i.e., Equation 3 in Appendix C) indicates both main effects to be significaist. Specifically, the Event-based/Predefined (κ=0.553, σ_κ=0.017) lead to significantly higher agreement between raters than Recently/Undefined (κ=0.423, σ_κ=0.019) time window (Z=5.00, p=.00); and the 4AFC (κ=0.421, σ_κ=0.019) lead to significantly lower agreement between raters than 3AFC (κ=0.507, σ_κ=0.020) response alternative (Z=3.16, p=.00). Figure 12 and Figure 13 illustrates the main effects of time windows and response alternatives, respectively.

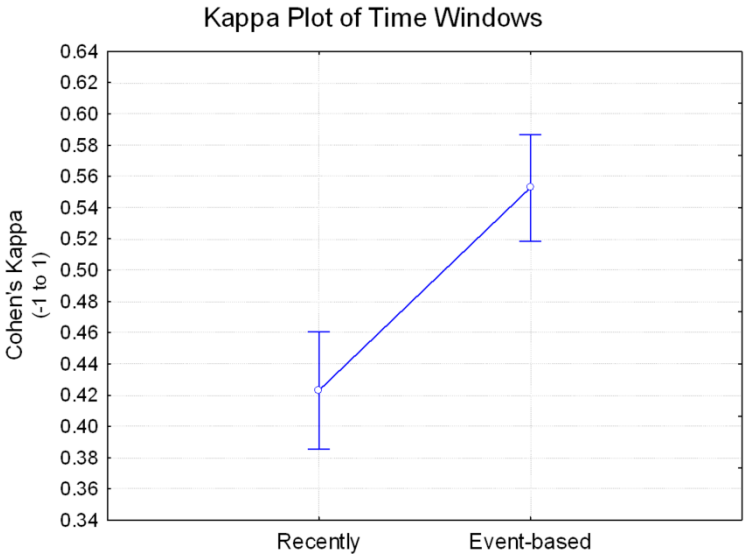


Figure 12: Plot of aggregated Cohen's Kappa for the main effect of Time Window.

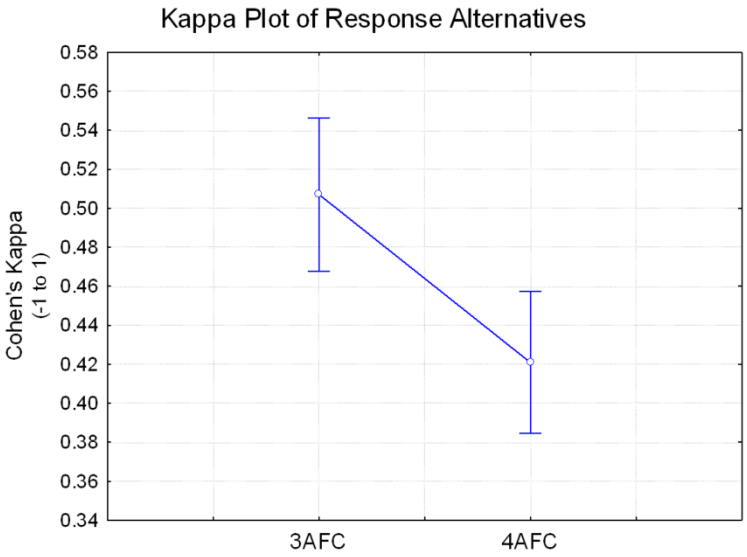


Figure 13: Plot of aggregated Cohen's Kappa for the main effect of Response Alternative.

In summary, the analysis with κ's led to two findings: (i) the Event-based improves reliability over Recently time windows (consistent with hypothesis), and (ii) 4AFC hinder reliability compared to 3AFC response alternatives (in contradiction with hypothesis).

The Process Overview Measure: Methodological Developments to Enhance Inter-Rater Reliability

Nathan Lau, Gyrd Skraaning jr, Tommy Karlsson, Christer Nihlwing, & Greg A. Jamieson

CEL 11-02

Cognitive Engineering Laboratory, Department of Mechanical & Industrial Engineering, University of Toronto
5 King's College Road, Toronto, Ontario, Canada M5S 3G8
URL: www.mie.utoronto.ca/labs/cel/

Directors: Kim J. Vicente, Ph.D., P. Eng.
Greg A. Jamieson, Ph D., P. Eng.

The Cognitive Engineering Laboratory (CEL) at the University of Toronto (U of T) is located in the Department of Mechanical & Industrial Engineering, and is one of three laboratories that comprise the Human Factors Research Group. CEL was founded in 1992 and is primarily concerned with conducting basic and applied research on how to introduce information technology into complex work environments.

Current CEL Research Topics

CEL has been funded by Atomic Energy Control Board of Canada, AECL Research, Alias|Wavefront, Asea Brown Boveri Corporate Research - Heidelberg, Canadian Foundation for Innovation, Defence Research & Development Canada (formerly Defense and Civil Institute for Environmental Medicine), Honeywell Technology Center, IBM, Japan Atomic Energy Research Institute, Microsoft Corporation, Natural Sciences and Engineering Research Council of Canada, Nortel Networks, Nova Chemicals, Westinghouse Science & Technology Center, and Wright-Patterson Air Force Base. CEL also has collaborations and close contacts with the Mitsubishi Heavy Industries and Toshiba Nuclear Energy Laboratory. Recent CEL projects include:

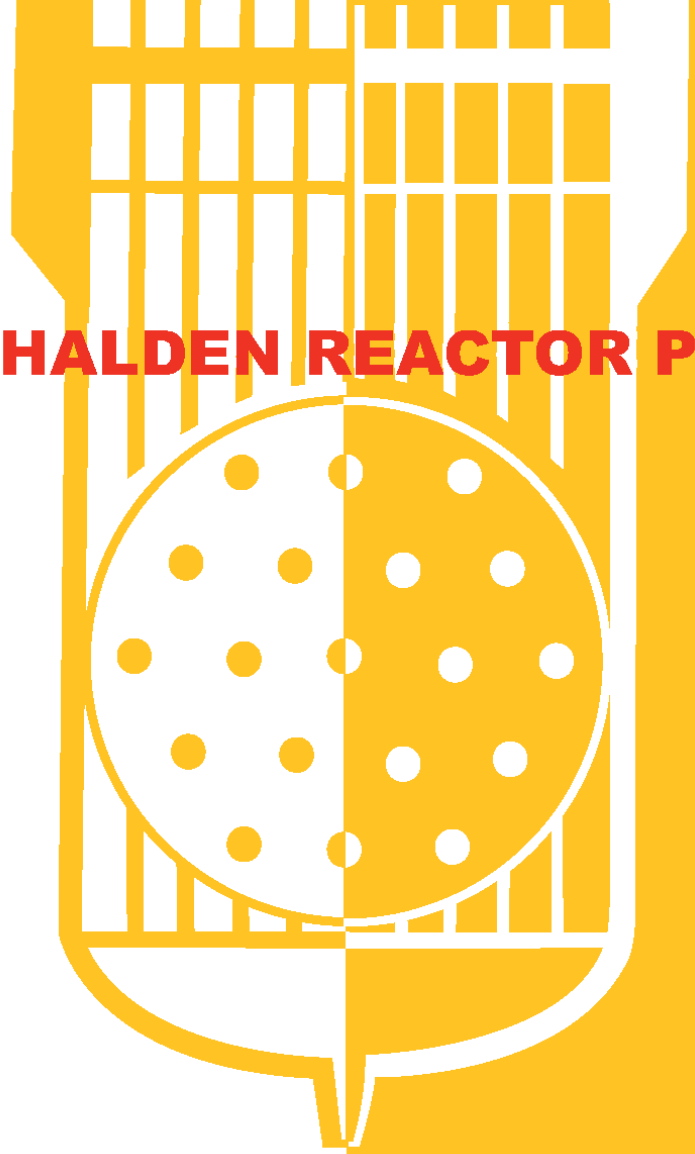
- Developing advanced human-computer interfaces for the petrochemical and nuclear industries to enhance plant safety and productivity.
- Understanding control strategy differences between people of various levels of expertise within the context of process control systems.
- Developing safer and more efficient interfaces for computer-based medical devices.
- Creating novel measures of human performance and adaptation that can be used in experimentation with interactive, real-time, dynamic systems.
- Investigating human-machine system coordination from a dynamical systems perspective.

CEL Technical Reports

For more information about CEL, CEL technical reports, or graduate school at the University of Toronto, please contact Dr. Kim J. Vicente or Dr. Greg A. Jamieson at the address printed on the front of this technical report.

HWR-971

OECD HALDEN REACTOR PROJECT



OECD

The Process Overview Measure: Methodological Developments to Enhance Inter-Rater Reliability

INSTITUTT FOR ENERGITEKNIKK
OECD HALDEN REACTOR PROJECT
P.O. BOX 173, NO-1751 HALDEN, NORWAY
www.ife.no/hrp

HWR-971

OECD HALDEN REACTOR PROJECT

The Process Overview Measure: Methodological Developments to Enhance Inter-Rater Reliability

by

Nathan Lau, Gyrd Skraaning Jr, Tommy Karlsson, Christer Nihlwing, OECD Halden Reactor Project; Greg A. Jamieson, University of Toronto

2011-02-09

FOREWORD

The experimental operation of the Halden Boiling Water Reactor and associated research programmes are sponsored through an international agreement by:

- the Institutt for energiteknikk (IFE), Norway,
- the Belgian Nuclear Research Centre SCK•CEN, acting also on behalf of other public or private organisations in Belgium,
- the Risø DTU National Laboratory for Sustainable Energy, Technical University of Denmark,
- the Finnish Ministry of Employment and the Economy (TYÖ),
- the Electricité de France (EDF),
- the Gesellschaft für Anlagen- und Reaktorsicherheit (GRS) mbH, representing a German group of companies working in agreement with the German Federal Ministry of Economics and Technology,
- the Japan Nuclear Energy Safety Organization (JNES),
- the Korean Atomic Energy Research Institute (KAERI), acting also on behalf of other public or private organisations in Korea,
- the Spanish Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), representing a group of national and industry organisations in Spain,
- the Swedish Radiation Safety Authority (SSM), representing public and privat nuclear organisations in Sweden,
- the Swiss Federal Nuclear Safety Inspectorate ENSI, representing also the Swiss nuclear utilities (Swissnuclear) and the Paul Scherrer Institute,
- the National Nuclear Laboratory (NNL), representing a group of nuclear licensing and industry organisations in the United Kingdom, and
- the United States Nuclear Regulatory Commission (USNRC),

and as associated parties:

- Japan Atomic Energy Agency (JAEA),
- the Central Research Institute of Electric Power Industry (CRIEPI), representing a group of nuclear research and industry organisations in Japan
- the Mitsubishi Nuclear Fuel Co., Ltd. (MNF)
- the Czech Nuclear Research Institute (NRI),
- the French Institut de Radioprotection et de Sûreté Nucléaire (IRSN),
- the Ulba Metallurgical Plant JSC in Kazakhstan,
- the Hungarian Academy of Sciences, KFKI Atomic Energy Research Institute,
- the JSC "TVEL" and NRC "Kurchatov Institute", Russia,
- All-Russian Research Institute for Nuclear Power Plants Operation (VNIIAES), Russia,
- the Slovakian VUJE - Nuclear Power Plant Research Institute, and
- EU JRC Institute for Transuranium Elements, Karlsruhe,

and associated parties from USA:

- the Westinghouse Electric Power Company, LLC (WEC),
- the Electric Power Research Institute (EPRI),
- the Global Nuclear Fuel (GNF) – Americas, LLC and GE-Hitachi Nuclear Energy, LLC, and
- the US Department of Energy (DOE)

The right to utilise information originating from the research work of the Halden Project is limited to persons and undertakings specifically given this right by one of these Project member organisations.

Recipients are invited to use information contained in this report to the discretion normally applied to research and development programmes. Recipients are urged to contact the Project for further and more recent information on programme items of special interest.

Institutt for energiteknikk
OECD HALDEN REACTOR PROJECT

Title

The Process Overview Measure: Methodological Developments to Enhance Inter-Rater Reliability

Author:

Nathan Lau, Gyrd Skraaning Jr, Tommy Karlsson, Christer Nihlwing, OECD Halden Reactor Project; Greg A. Jamieson, University of Toronto

Document ID:
HWR-971

Keywords:
Situation Awareness, Process Overview, Measure, Human Factors, Inter-rater Reliability

Abstract:

The Process Overview Measure assesses monitoring performance by eliciting knowledge about parameter behaviours using the freeze-and-query techniques. Participants from an earlier empirical study reported ambiguities associated with time references in the queries and completeness of the response alternatives, hindering the inter-rater reliability of the measure. An empirical study was conducted to evaluate two methods of improving the inter-rater reliability of the Process Overview Measure. The first method aimed to reduce ambiguities in the time reference by specifying scenarios events to define the time windows for judging parameter behaviours. The second method aimed to improve the completeness of the response alternatives by adding "fluctuating around the same point" as a category of parameter behaviour in the response set. The results indicate that specifying scenario events to define time windows in the queries could improve inter-rater reliability. However, adding "fluctuations" in the response set hindered inter-rater reliability. The results support the use of scenario events to define time windows for judging parameter behaviours in the queries. A revised procedure to the Process Overview Measure is presented.

Issue Date:		Name	Signature	Date
2011-02-09	Prepared by:	Nathan Lau	Sign.	2010-11-30
Confidential grade:	Reviewed by:	Greg A. Jamieson	Sign.	2010-11-30
HRP Only	Approved by:	A. Bye	Sign.	2011-02-09

MAIL	TELEPHONE	TELEFAX		TELEFAX	
P.O. BOX 173	+47 69 21 22 00	Administration	+ 47 69 21 22 01	IND/OC div.	+ 47 69 21 24 90
NO-1751 HALDEN		Nuclear Safety	+ 47 69 21 22 01	VISIT/RID/COSS div.	+ 47 69 21 24 60
Norway		Purchasing Office	+ 47 69 21 24 40	Reactor Plant	+ 47 69 21 24 70

TABLE OF CONTENTS

1. INTRODUCTION

As the operation of nuclear power plants (NPPs) becomes increasing knowledge-based (e.g., see [1]), the notion of Situation Awareness (SA), along with situation assessment, becomes increasingly useful for describing and discussing operator work [2]. However, the application of SA in nuclear process control faces two fundamental challenges. First, SA is poorly defined or, at least, suffers from the lack of a unanimous definition or model as the ambitious scope of the notion aims to cover a vast variety of cognitive work across multiple domains. Second, compounding the lack of consensus about the notion, there is little SA research specific to nuclear process control (or process control in general), especially by comparison to other domains such as aviation [3]. Consequently, the nuclear domain cannot fully capitalize on the SA notion. The OECD Halden Reactor Project (HRP) is addressing the challenges in the application of SA by developing domain-specific characterizations and measures. [4] conceptualizes SA in nuclear process control as consisting of three components – Process Overview, Scenario Understanding, and Metacognitive Accuracy. These components reflect awareness/knowledge acquired through monitoring, diagnosis and self-assessment activities, respectively. [4] further presents measures corresponding to the three SA components. HRP continues to advance SA research in nuclear process control with further development of Process Overview. [5] develops Process Overview by extending the descriptions of the interactions between domain characteristics and monitoring behaviours, and assesses the Process Overview Measure by conducting three empirical studies.

Improving the Process Overview Measure could lead to more accurate and reliable indicator of monitoring performance that would yield better evaluation of control room designs to support operators. This report continues the development of the Process Overview Measure. Specifically, it documents an empirical study evaluating two methodological modifications to the Process Overview Measure as guided by the empirical findings in [5]. The report begins with a review of the Process Overview Measure followed by the experimental method, analysis and results, and discussion of the empirical findings.

2. THE PROCESS OVERVIEW MEASURE

The Process Overview Measure is a domain-specific SA measure that aims to assess operator knowledge/awareness acquired through monitoring. It is developed in accordance with Process Overview – a domain-specific characterization of SA and adapts the Situation Awareness Control Room Inventory (SACRI) [6, 7] and Situation Awareness Global Assessment Technique (SAGAT) [8, 9]. This chapter describes the Process Overview Measure, reviews the current research on the measure, and identifies corollary issues that ultimately lead to the empirical study presented in the subsequent chapters. (Refer to [5] for the full formulation of the Process Overview concept and measure.)

2.1 Description

The Process Overview Measure operationalizes Process Overview as the accurate detection of meaningful changes in relevant process parameters. Process parameters are *relevant* when they effectively represent the operating contexts (e.g., shutdown) and reveal potential process anomalies. Parameter changes are *meaningful* when they represent the systematic trends as opposed to (uninformative) fluctuations.

The Process Overview Measure involves three phases: preparation, data collection and scoring. During the preparation phase of full-scope simulator experiment or evaluation session, process

experts are instructed to perform three inter-connected tasks – development/review of scenarios, selection of relevant parameters, and identification of simulator-freeze points for query administration. These tasks are described in the following paragraphs.

Process experts are typically responsible for developing test scenarios with the characteristics that are useful for the purpose of the study (see [10] for a discussion.) For instance, to evaluate an alarm system for monitoring, the test scenarios must contain process events or faults leading to alarms. The Process Overview Measure does not prescribe any guidance for developing scenarios because the dominant consideration should be the purpose of the empirical study. However, the Process Overview Measure does rely on process experts to develop representative test cases that are relevant to experimental topics and sufficiently challenging to operators.

After the development of the scenarios, process experts select process parameters according to the scenario characteristics. Relying on their knowledge in developing the scenarios, process experts should select a set of *relevant* process parameters that represents the operating context and process events (including faults) in the scenario. In other words, the operators/participants successfully completing the scenarios are expected to know the behaviours of these parameters while monitoring the process. The awareness of these parameter behaviours is elicited through administrations of queries in the form as specified in Figure 1. For empirical studies where process experts do not develop test scenarios but still implement the Process Overview Measure, they must review the scenarios carefully.

Process Overview Query Structure:
Recently, the parameter [code] has:

Process Overview Response Alternatives:
Increased/Stayed the same/Decreased

Figure 1: Query and response format of the Process Overview Measure.

Relevant parameters can typically be classified as (i) context-sensitive or (ii) fault-sensitive according to Process Overview [5]. Context-sensitive parameters reflect the overall plant states based on the operating contexts given to the operators at the beginning of the scenarios. For instance, during start-up at a certain power level, operators often sample a set of key parameters periodically to determine the general progress. The cuing effects for context-sensitive queries should be negligible as these parameters are emphasized during their professional training and work practice. Fault-sensitive parameters reveal the process faults introduced by the scenarios. Therefore, fault-sensitive parameters require close observation. Hence, operators may not sample these fault-dependent parameters during normal operations. Thus, fault-sensitive queries may be subject to cuing effects, prompting consideration of the method by which these queries are introduced (see below). Note that the two classes of parameters could overlap.

Process experts also select timing of simulator freezes in the scenarios to administer queries about the relevant parameters. The timing of these freezes (i.e., administration of the queries) should be determined according to selection of process parameters and scenario characteristics. Some context-sensitive parameters become relevant or irrelevant as the scenarios progress. Context-sensitive queries that are relevant for the entire scenarios may be administered at random times. Fault-sensitive queries often require strategic timing of freezes because the queries need to coincide with the

introduction of the faults without resulting in cuing effects[1]. Two general methods are available to counteract cuing effects of administering fault-sensitive queries. The first method relies on the strategic timing of alarms that occur immediately after the freeze to nullify cuing across participants. The second method relies on administering the queries at the end of the scenarios when the cues from the queries cannot influence operator performance. Either or both method may be employed in any implementation of the Process Overview Measure.

The three tasks performed by process experts in preparation for an experiment - scenario development/review, parameter selection, and freeze identification – are inter-connected and not necessary sequential. For instance, the scenarios may be re-designed to provide effective strategic timing of freezes that can nullify cuing effects. Flexibility across those three tasks should be leveraged to optimise the quality of the SA measurements with respect to the purpose of the empirical studies.

During data collection (i.e., while operators/participants are running the test scenarios), the simulator should freeze according to the timings specified by the process experts during data preparation phase. During simulator freezes, the participants answer the corresponding set of queries without any access to process displays. From here onwards, the participant answers are labelled as “responses”. At the same time, the process experts supporting the data collection answer the queries with access to all the process displays. From here onwards, the process expert answers are labelled as “reference keys”. In addition to collecting the responses and reference keys to the queries, the simulator should log the parameters throughout the scenario for potential verification needs after the experiment.

After the data collection, final scores are the proportion correct (or matches) between the responses and reference keys (collected from the participants and process experts, respectively). These scores may then be analyzed statistically.

2.2 Current research and corollary issues

2.2.1 Empirical Findings on the Process Overview Measure

Current research[2] indicates that the Process Overview Measure is *sensitive* to experimental manipulations [5]. In a Haden Man-Machine LABoratory (HAMMLAB) experiment, Process Overview Measure illustrates the relative advantage between three display types for monitoring NPPs under different operating situations. In another full-scope simulator experiment intended to explore futuristic operational concepts [5], the Process Overview Measure illustrated the impact of transparent automation displays with respect to different staffing configurations on operator monitoring performance. Both experiments also produced other empirical results that corroborated with the Process Overview findings. In summary, the Process Overview Measure revealed the impact of new operational concepts or technology on monitoring performance.

Besides sensitivity, the Process Overview Measure demonstrated acceptable inter-rater *reliability* between process experts. In an empirical study comparing the references keys between three process experts [5], the Process Overview Measure exhibited "fair" to "substantial" agreement (i.e., Kappa between 0.4-0.6), which is reasonable for judgment tasks involving complex information processing (e.g., medical diagnosis as suggested by [11]) as in monitoring NPPs. Note that slightly higher agreement (i.e., Kappa between 0.7-0.8) would be desirable. Two sources of ambiguity are discovered

[1] Cueing effect refers to a shift in participant attention (consciously or unconsciously) due to the appearances or presence of some environmental stimuli. Typically, this effect is unintended in the experiments.

[2] For the purpose of this report, the literature review only focuses on research specific to the Process Overview Measure. General discussion on SA measurement is widely available in the literature (e.g., [26-28]). The literature also contains some discussion on SACRI [6, 7] and SAGAT [8, 9] that inspired the development of the Process Overview Measure.

from discussion with operators and process experts (who provided the responses and reference keys in the empirical studies) during the debriefing sessions. First, the term "Recently" in the Process Overview queries is deemed ambiguous as operators can think of multiple events during the scenario that can be considered important for defining the relevant time period to judge parameter changes. Second, some parameters fluctuate up and down during the scenarios, which are not interpreted as strictly increasing, staying the same or decreasing. These comments prompt a search for methodological improvements in the Process Overview Measure.

2.2.2 Methodological Development

The qualitative and quantitative results from the empirical studies [5] direct research focus towards specifying (i) a clear time reference in the queries, and (ii) comprehensive categories of parameter behaviours in the response set.

To provide a clear time reference in the queries, a potential solution is to specify the starting point of a critical scenario event as the starting time for determining parameter changes as opposed to leaving the judgment (of time) entirely to the operators. Specifying scenario events is more consistent than a time (i.e., number of minutes) with Process Overview [4, 5], which indicates that process operators think in action time, than specifying absolute/clock time in the query [12].

To capture all relevant categories of parameter behaviours, the response set used in [4, 13, 14] requires augmenting with a “fluctuation” option. While the current response set may be conceptually complete from the perspective that any parameter changes must fit into one of the three options (i.e., decreased, stayed the same, and increased), process operators may think in more categories of parameter behaviours that offer additional operational utilities in practice. That is, a parameter may fluctuate abnormally around a value signalling process anomalies. In such cases, even if all operators and process experts know that a parameter is fluctuating, they must select another option given the current Process Overview response set. Some may consider the parameter unchanged while others may judge the parameter to have increased or decreased depending on the exact parameter value at the time of the simulator freeze.

2.3 Objective of the this study

To assess the merits of specifying scenario events for time references and adding fluctuations to the response set, we conducted a controlled experiment to compare the inter-rater agreements between four different variants of the Process Overview Measure.

3. METHOD

3.1 Experimental Design

Each participant performed 56 trials, comprised of fourteen different cases assigned to four treatment combinations (14x4). In total, the data set contained 168 trials from three raters (56x3). The three raters also yielded three combinations of rater-pairs (i.e., Rater A & B, A & C, and B & C) for calculating agreement between any two experts (also see section 4.1).

A 3x2x2 split-plot factorial (SPF) design was employed with a between-subject factor of rater-pairs (AB, AC, and BC), and within-subject factors of time window (Recently/Undefined and Event-based/Predefined) and response alternative (3AFC and 4AFC).

The treatments were completely crossed and counterbalanced by randomizing presentation order of cases and treatments with two restrictions. The first restriction was that each treatment combination occurred only once for each case. The second restriction was that the fourteen different cases were randomly selected without replacement until all cases were selected.

3.2 Participants

The participants were three male process experts. Besides accessibility, one principal selection criteria was their familiarity with two previous empirical studies that investigated the Process Overview Measure. As mentioned in *Procedure and experimental task* (section 3.4), one experiment employing the HAMBO full-scope simulator provided the source data generating the parameter trends. The scenario trials in which participants operated the HAMBO simulator provided the corresponding operating contexts of those parameter behaviours in this study. The second empirical study investigated the inter-rater reliability of the Process Overview Measure that also relied on the full-scope simulator study for its source data. All three process experts participated in the two earlier studies (i.e., Study 2 and 3 in [5]), providing them with the exposure to the simulator, the Process Overview Measure, and scenarios in full-scope simulator experiment. The exposure to the two related studies should prepare these process experts (better than others) to mentally simulate the parameter behaviours with respect to the operating contexts (i.e., scenario descriptions and operator actions given on papers) when they could not observe the development of the nuclear process in real time (i.e., during the full scope experiment). However, the selective recruitment process would mean that "rater" is a fixed as opposed to a random factor (i.e., participant is typically a random factor in both full-scope simulator and inter-reliability studies), limiting the generalization of the results on this particular factor[3].

The relevant characteristics of each process expert are summarized below:

Rater A was a process expert employed at the HRP (for two years). He worked as a control room operator in multiple nuclear plants for fifteen years and participated as a process expert to develop of the HAMBO simulator in late 1990's. For the experiment where the source data for the graphs was drawn in this study, he designed all of the scenarios and advised other HRP scientists on the processes and operations of the physical and simulator plant. He also identified all of the process parameters for the Process Overview queries. During data collection, Rater A played the role of field operator and electrician as required by the scenarios and participant interventions. He was also the designated process expert for completing questionnaires related to participant performance.

Rater B is a recently retired (for around one year) shift supervisor at the nuclear that the HAMBO simulator replicates. He worked as a control room shift supervisor for 32 years and participated in about ten HRP studies prior to this study. He was a participant for the pilot trials of the experiment of the source data, providing additional exposure to the experiment. Furthermore, Rater B acted as the second expert rater (in addition to Rater A) for investigating inter-rater reliability of the Process Overview Measure during the data collection of the full-scope simulator experiment.

Rater C was a process expert employed at the HRP for eleven years. He worked as a control room operator in multiple nuclear plants for fifteen years and participated as a process expert in the development of the HAMBO simulator. Prior to this study, he had similar responsibility for numerous HAMMLAB simulator studies. For the experiment where the source data for the graphs was drawn in

[3] There is no practical solution to the trade off between selective recruitment and generalizablity. Randomly recruited process experts would hinder the validity of the test due to lack of knowledge on the experiment to judge parameter changes. On the other hand, selectively recruited process experts impose some limits on generalization of the statistical results.

this study, he developed the scripts that automatically start-up the simulator plant, implemented the scenarios, co-designed the transparent automation interface and advised other HRP scientists on the processes and operations of the physical and simulator plant. During data collection, he was the designated technical support person for simulator operations. For the earlier inter-rater reliability study, Rater C answered the Process Overview queries after the full-scope simulator experiment (in a set up similar to the treatment combination of Recently/Undefined and 3AFC treatment combination of this study). He responded to queries based on descriptions of scenarios and operator actions as well as parameter trends.

3.3 Experimental manipulations

This experiment included two experimental manipulations – time windows and response alternatives of the Process Overview Measure queries.

3.3.1 Time windows of Process Overview Measure queries

Time window refers to how Process Overview Measure queries frame the time periods for the operators to judge the behaviours of process parameters. This experiment included two types of time windows – (i) Recently/Undefined, and (ii) Event-based/Predefined.

The Recently/Undefined time window (a method employed prior to this study [5]) relied on operators to decide on the time period by using the term “Recently” for judging parameter behaviours (see Figure 1). The operators were instructed to define the start time of the period at the last meaningful plant state and the end time at the simulator freeze for the particular parameter/query.

The Event-based/Predefined time window relied on process experts to predefine the time period by specifying events in the scenarios (during scenario development) for judging parameter behaviours. The query defined the start time at the occurrence of a specific event in the scenario and end time at the simulator freeze to be the period for judging parameter behaviours (see Figure 2).

Since event X (e.g., telephone call from field operators) until now, parameter [code] (e.g., condenser pressure) has:

Figure 2: Process Overview query structure with the Event-based/Predefined time window.

3.3.2 Response alternatives of Process Overview queries

The response alternative refers to the categories of parameter behaviours presented to the operators for answering the Process Overview queries. This experiment included two sets of response alternatives – (i) 3 alternative-forced choice (3AFC) and (ii) 4 alternative-forced choice (4AFC).

The 3AFC response set (a method employed prior to this study) included three categories of parameter behaviours: (i) increased, (ii) stayed the same, and (iii) decreased. The 4AFC response set included four categories: (i) increased, (ii) stayed the same, (iii) fluctuated around the same point, and (iv) decreased.

3.4 Procedure and experimental task

The data collection involved three stages: (i) a 0.5-hour session of introduction and instruction, (ii) eight 1.25-hour sessions of experimental task, and (iii) a 0.5-hour debriefing session.

3.4.1 Introduction and instruction

During the introduction and instruction session, the experimenter verbally described the intent, environment and schedule of the data collection for the study. The instruction of the experimental tasks was delivered in both verbal and written form (Appendix A). The participants were instructed to answer Process Overview queries by inspecting trend graphs generated from logs of a prior full-scope simulator study [5, 13, 14] and considering contexts in which the parameter trends were being developed.

The participants were informed that fourteen cases (i.e., the contexts) were drawn from the prior full-scope simulator study (in which they had participated as process experts). The full-scope simulator experiment collected data from nine NPP crews operating under eight different scenarios (see [13, 14] for descriptions). Eight and six scenarios performed by the second last and the last crews, respectively, made up the fourteen cases in this study. These fourteen cases were selected because comparisons could be made with the dataset collected from a prior reliability study [5] on the Process Overview Measure, which relied on the last two crews of the full-scope simulator study.

Because participants were not situated in the context as the parameters changed (i.e., real time during the full-scope simulator experiment), the instruction explicitly prompted them to consider contextual information as well as scales of the axes when answering Process Overview queries. The Process Overview queries were selected by process experts based on scenario analyses to reflect the expected knowledge from monitoring the nuclear process in those specific scenarios. The explicit reminder was aimed to encourage participants to rely on their expertise in the nuclear process and knowledge of the full-scope experiment that would direct monitoring behaviours during the actual scenario trials. The contextual information was provided on paper through scenario descriptions, task performance scores and brief descriptions of the operator actions with respect to the fourteen cases. The experimenter also informed the participant that they could request the trend graphs to be generated on different scales that deemed appropriate for the parameters in the corresponding contexts. When they were uncertain about the exact sensor of the parameter code, the participant could request access to the simulator in “freeze” mode to identify the sensors pertaining to the specific parameter codes.

After reviewing the instruction, the participants were asked to answer four practice queries corresponding to the four different experimental conditions as described in the section on Experimental task below. Before the data collection sessions began, the participants were asked whether any clarification was necessary.

3.4.2 Experimental task

Data from each participant were collected over eight sessions. Each session took approximately one and a quarter hour, separated by breaks of at least fifteen minutes to avoid fatigue. Each session contained seven cases/trials, each of which contained fourteen to eighteen Process Overview queries[4].

[4]The number of Process Overview queries varies across cases/trials due to characteristics of the scenarios for those cases.

For the treatment of the time window with the level of Recently/Undefined, the participants were required to (i) read the query, (ii) review the graph, (iii) explicitly define the time window by shading in a portion of the graph (with a mouse extending a rectangle; Figure 3), and (iv) select one of the response alternatives with a mouse).

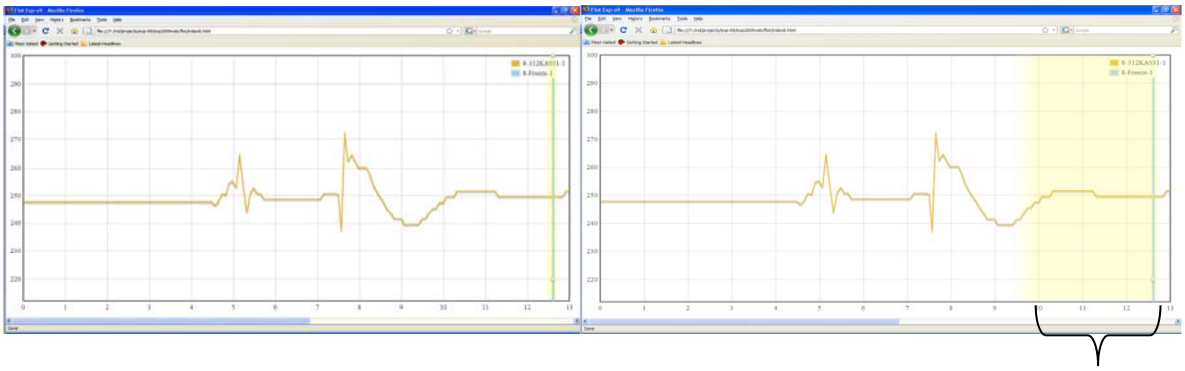


Figure 3: Defining time windows for the Recently/Undefined time window condition.

For the treatment of the time window with the level of Event-based/Predefined, the participants were required to (i) read the query, (ii) review the graph (Figure 4), and (iii) select one of the response alternatives with a mouse. This treatment level did not require operators to shade in any time period for judgment.

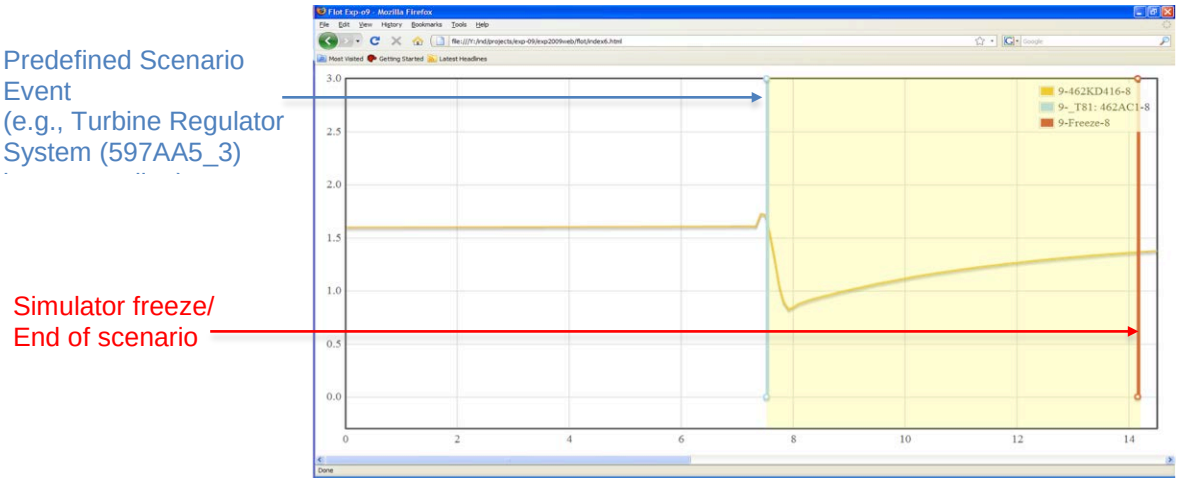


Figure 4: Graph for the Event-based/Predefined time window condition.

For the treatment of response alternative, the 3AFC differed from the 4AFC only in that the response alternative of “fluctuating around the same point” was available for the 4AFC but not the 3 AFC conditions. The two treatments (i.e., experimental manipulations) consisted of two levels each, resulting in four different combinations of experimental manipulations.

Well-defined criteria applicable to all Process Overview queries did not exist for identifying time windows and selecting response alternatives. The participants were explicitly instructed to select a response to a query that would be most meaningful for classifying the parameter behaviour given the specific operating context (i.e., the case). For instance, the response alternative of “fluctuating around the same point” should only be selected if knowledge of the fluctuation offered value to control room operators in operating the NPPs. Small fluctuations with no operational significance should be considered noise and therefore classified as “stay the same”.

3.4.3 Debriefing

After all eight sessions, the data collection ended with a debriefing session to discuss the experimental manipulations in a semi-structured interview format (Appendix B). The debriefing permitted the participants to express their preference and potential improvements to the Process Overview Measure.

3.5 Experimental environment

The data was collected in a large office with two computers, two 30" LCD displays and two 22" LCD displays. One computer was used to present the graphs in PowerPoint files on one connected 30" displays and to collect responses of the participants through the Halden Questionnaire Systems [15] with the other connected to a 30" display. The 22" LCD displays were inactive during the experiments. The second computer was running the HAMBO simulator in "freeze" mode in the background. It was accessible by changing the input signal on one of the 30" LCD displays. The second computer was available in case the participants needed help in recalling the parameter codes with respect to the sensors. (None of participants asked to use the simulator for recalling parameters.) In addition, the participants were provided with all the scenario descriptions and task performance scores with brief descriptions of the operator actions on papers corresponding to the fourteen cases.

3.6 Hypothesis

The two proposed modifications to the Process Overview Measure were expected to improve consistency between raters. First, specifying a scenario event, at which considerations of the parameter behaviours should start, could reduce the ambiguities in comparison to the term "recently". Second, the 4AFC response set could provide a qualitatively distinct option suitable to the behaviours of some parameters thereby increasing agreement between raters over the 3AFC response set (when marginal distribution/chance agreement is factored in).

In addition to hypotheses related to the modifications, the level of agreement between raters could vary based on empirical the findings of an earlier reliability study [5].

4. ANALYSIS AND RESULTS

4.1 Statistical testing

The data analysis involved a two-step statistical procedure (i) calculating Cohen's Kappa, κ (see [16, 17] and Appendix C for a brief review), between raters/participants for each trial[5], and (ii) performing significance testing of differences between κ scores between rater pairs using ANOVA and non-parametric tests. (Note that the κ statistics account for the reduced chance agreement with the increased number of response options.)

The two-step analysis procedure should yield the most accurate results of the collected data. The first step of calculating κ resulted in statistics of agreement between two raters that accounted for the marginal distribution (i.e., chance agreements and number of response options). As mentioned, three

[5] The adopted aggregation procedure, calculating κ between raters per trial, is not ideal as calculating κ between raters per experimental condition most likely produces the best estimates. Furthermore, ANOVA is not the typical significant testing method for κ's (see Appendix C). However, aggregation over experimental conditions (as opposed to trials) limits analysis to graphical inspection and pair-wise comparisons, failing to provide omnibus tests of the full experimental design. Appendix D documents the analysis of κ's based on data aggregation per experimental condition relying on graphical inspection and pair-wise comparisons.

participants led to three rater-pairs and thus three set of κ's (i.e., Rater A & B, A & C, and B & C). The total number of κ's was 168 (3 rater-pairs x 2 levels of time-windows x 2 levels of response alternative x 14 cases). The second step applied ANOVA on κ's to perform significance testing on the full experimental design.

Two Process Overview queries were omitted from data analysis (one in the Recently and 3AFC, and another in the Recently and 4AFC experimental conditions) because of an error with the two graphs. As a result, both 3AFC and 4 AFC in the Recently time window condition contained 225 data points as opposed to 226 data points in the Event-based time window condition.

The κ's were analyzed with ANOVA using Type III/Unique sums of squares with a between-subject factor of rater-pairs (AB, AC, and BC), and within-subject factors of time window (Recently/Undefined and Event-based/Predefined) and response alternative (3AFC and 4AFC).

The multivariate normality assumption was not satisfied. Two of the twelve distributions were not normally distributed. Specifically, the distribution of κ between Rater A and B in the condition of Event-based time window and 4AFC response alternative was negatively skewed; whereas, the distribution of κ between Rater B and C in the condition of Recently time window and 3AFC response alternative was positively skewed. Though generally robust to violations of the normality assumption, ANOVA results would not be particularly robust to distributions heterogeneous in form [18]. Levene's test indicated homogeneity of variance, but there was a *slight* positive correlation between means and standard deviations for the between factors of rater-pairs that could threaten the robustness of the ANOVA results [19]. The sphericity assumption was inherently satisfied as all within-subject treatments only consisted of two levels. From the perspective of the entire data set, the aforementioned characteristics of the raw data were not "gross" violations of ANOVA assumptions. Therefore, the ANOVA results on κ statistics are presented. To be cautious, non-parametric tests [20-22] were employed to verify the subset of the experimental design only containing the significant ANOVA results.

The ANOVA on κ (Table 1) revealed: (a) between-subject main effect of rater-pairs ($F(2,39)=7.85$, $p=0.00$, $\eta^2=.29$, Figure 5); (b) within-subject main effect of time window ($F(1,39)=26.29$, $p=0.00$, $\eta^2=.40$, Figure 6); (c) within-subject main effect of response alternative ($F(1,39)=12.04$, $p=0.00$, $\eta^2=.25$, Figure 8); and (d) interaction effect of rater-pairs and time window ($F(2,39)=3.23$, $p=0.05$, $\eta^2=.14$, Figure 7).

Table 1: ANOVA of Kappa

	SS	Df	MS	F	p	η^2
Fixed, *Between* Effect						
Rater-pair	1.4823	2	0.7412	7.8459	0.00	0.29
Error	3.6841	39	0.0945			
Fixed, *Within* Effects						
Time Window	0.8714	1	0.8714	26.2923	0.00	0.40
Time Window*Rater-pair	0.2140	2	0.1070	3.2280	0.05	0.14
Error	1.2926	39	0.0331			
Response Alternative	0.2714	1	0.2714	13.0350	0.00	0.25
Response Alternative*Rater-pair	0.0853	2	0.0426	2.0479	0.14	0.10
Error	0.8121	39	0.0208			
Time*Window*Response Alternative	0.0101	1	0.0101	0.3821	0.54	0.01
Time*Window*Response Alternative* Rater-pair	0.0301	2	0.0151	0.5690	0.57	0.03
Error	1.0321	39	0.0265			

A series of four non-parametric tests, following the Holm's adjustment procedure of family-wise error rate [20], was performed to verify the significant ANOVA effects. The series of non-parametric tests confirmed all significant ANOVA effects:

- a non-parametric, Kruskal-Wallis One-way ANOVA for the main (between-subject) effect of rater-pair (H(2, N=168)=27.86, p=.00);
- a non-parametric, Wilcoxon Matched Pairs test for the main (within-subject) effect of time window (*Z*=4.93, N=84, p=.00, ES=.54)[6];
- a non-parametric, Kruskal-Wallis One-way ANOVA of rater-pair on the difference scores between Recently and Event-based conditions[7] for the interaction effect of rater-pair and time window (H(2, N=84)=7.74, p=.02); and
- a non-parametric, Wilcoxon Matched Pairs test) for the main (within-subject) effect of response alternative (*Z*=3.08, N=84, p=.00, ES=.34).

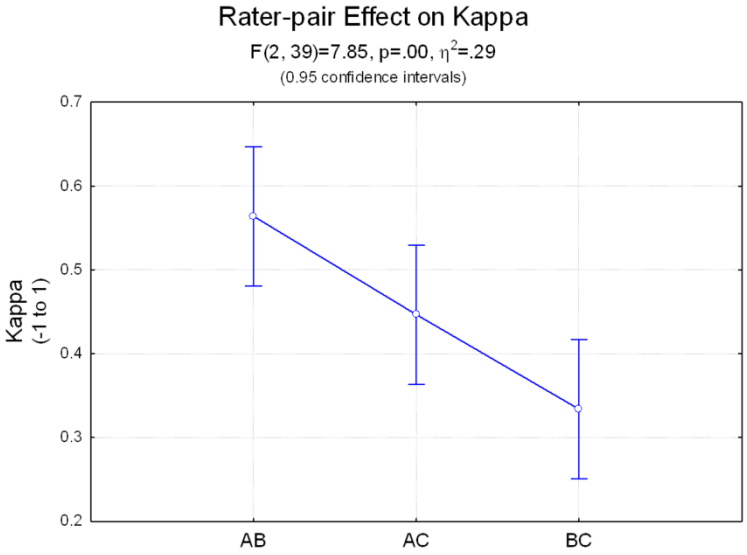
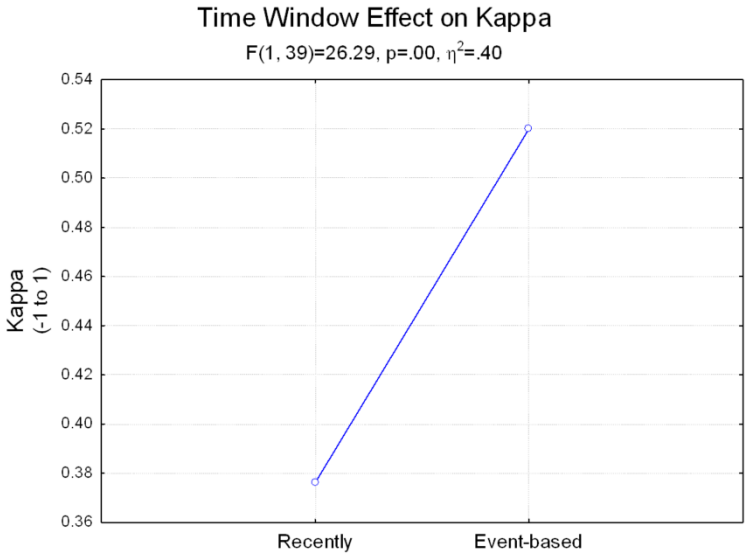


Figure 5: Rater-pair effect on Kappa.

Figure 5 illustrates that raters could bring unique influence in judging parameter changes as discovered in a prior reliability study on the Process Overview Measure [5]. In the prior study, the agreement was highest between raters most involved in the preparation of full-scope simulator study from where the source data was drawn (i.e., Rater A and C). In contrast, the highest agreement in this study was between Rater A and Rater B, who was less exposed to the full-scope simulator experiment than Rater A and C. Figure 7 illustrates the interaction effect between rater-pair and time windows that more precisely describes the influence of raters.

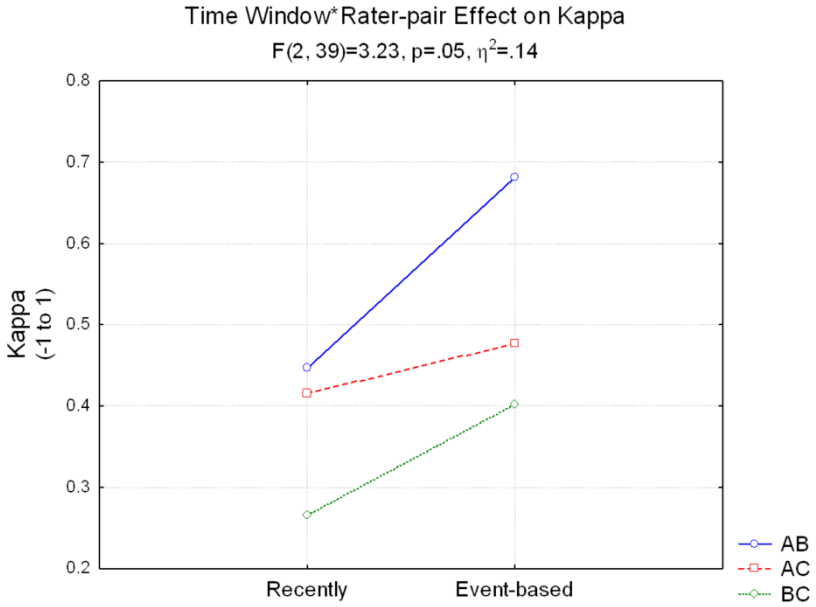
[6] ES denotes effect size for the Wilcoxon Matched Pairs test [22].

[7] As a non-parametric procedure was not available to test interaction effects, the difference scores between Recently and Event-based conditions were calculated, followed by a non-parametric test on rater-pair based on the difference scores. This method of verifying interaction effects is not a standard non-parametric statistical procedure found in the literature. The authors adapt some common statistical techniques to verify the ANOVA effects. This adapted procedure only works for two-way interaction effects in which one of the treatments only has two levels.



Time Windows	Mean	Std Dev.	95% Lower Bound CI	95% Upper Bound CI
Recently/Undefined	0.376	0.027	0.321	0.432
Event-based/ Predefined	0.520	0.028	0.464	0.576

Figure 6: Time window effect on Kappa.

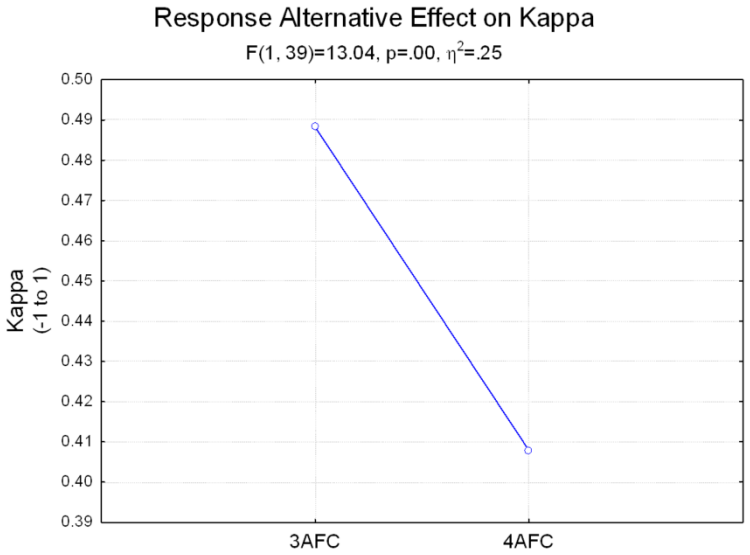


Rater-pairs	Time Windows	Mean	Std Dev.	95% Lower Bound CI	95% Upper Bound CI
AB	Recently/Undefined	0.446	0.048	0.350	0.543
AB	Event-based/ Predefined	0.681	0.048	0.585	0.778
AC	Recently/Undefined	0.416	0.048	0.320	0.512
AC	Event-based/ Predefined	0.477	0.048	0.380	0.574
BC	Recently/Undefined	0.266	0.048	0.169	0.362
BC	Event-based/ Predefined	0.402	0.048	0.305	0.499

Figure 7: Rater-pair and time window effect on Kappa.

Figure 6 illustrates that Recently time windows produced inferior agreement between raters compared to the Event-based ones. Figure 7 further illustrates that all pairs of raters appear to benefit with the

Event-based time windows but to varying degrees. The means of the rater-pairs in the Recently condition resembled the condition of the prior reliability study more than the Event-based condition. Furthermore, the agreement between pairs of raters in the Recently condition also more closely resembled the results of the earlier reliability study than suggested by the main effect of rater-pairs. Specifically, the difference in agreement between Raters A and B compared to Raters A and C was negligible while the agreement between Rater B and C was still the lowest. With respect to the experimental manipulation, Figure 5-Figure 7 illustrates that Event-based time windows could improve agreement between raters as stated in the hypothesis.



Response Alternative	Mean	Std Dev.	95% Lower Bound CI	95% Upper Bound CI
3AFC	0.488	0.026	0.435	0.541
4AFC	0.408	0.026	0.355	0.461

Figure 8: Response alternative effect on Kappa.

Figure 8 illustrates that 3AFC response alternatives produce superior agreement compared to the 4AFC ones in contradiction with the hypothesis.

4.2 Exploratory data analysis, observations and debriefing

The data collected in this study affords additional exploratory data analyses and some conjectural results. The exploratory analyses here involve substantial judgment of the analyst or investigate topics without sufficient experimental control. Although these results must be interpreted with caution, they shed light on the properties of the Process Overview Measure and may guide future research directions.

4.2.1 Judgment in Recently/Undefined time window

For the Recently/Undefined time windows, the participants were required to highlight the area of the trend graphs that defined their windows for judging parameter changes. The analyst coded these data in two ways: (i) the agreement of time windows between raters (i.e., process experts) on a binary scale, and (ii) the starting point of the self-defined time window at 0.5-minute precision. The quantification of these time windows inherently included some subjectivity of the analyst but allowed for a detailed investigation into the Recently/Undefined time window treatment level. Specifically, these

two analyses offer a coarse comparison between inter- and intra-rater consistency in defining Recently/Undefined time windows.

The agreement of the time windows between raters permitted an estimate of inter-rater agreement on defining “Recently”. The proportion agreement on the time windows in the Recently condition between raters (Table 2) was rather low, confirming the findings from the statistical testing above. Note that variation between the starting times as defined by different raters should **not** strictly be interpreted as a continuous scale. For instance, a three-minute difference in start time between two raters may have the same interpretation as a ten-minute difference because both cases suggest that the raters employ different criteria. For this reason, agreement of the time windows between raters as judged by experimenter could be more valid than correlation statistics of time window data quantified at 0.5-minute precision.

Table 2: Proportion agreement on defining time windows in the Recently treatment level.

Rater-pair	Proportion agreement on Recently time-windows
AB	.24
AC	.45
BC	.12

The start times of the windows as defined by the raters permitted another estimate of inter-rater agreement on defining “Recently”. Because time data is on a continuous scale, the intra-class correlations (ICC) statistics [23] are calculated as an estimate of inter-rater agreement on “recently” amongst all three raters. The ICC statistics (ICC(2,1)=.10, ICC(3,1)=.24)[8] are very low, confirming the low agreement between raters in defining “Recently” in the Process Overview queries.

The start times of the windows also permitted estimates of the internal consistency of the raters between the 3AFC and 4AFC response alternative in the Recently/Undefined time window condition. For a continuous scale of time data, ICCs are calculated as estimates of internal consistency[9]. ICC(3,1) for Rater A, B, and C are .51, .41, and .56, respectively. The correlations indicate low to moderate internal consistency within raters in defining “Recently”.

As mentioned, variation in the starting times should not strictly be interpreted as a continuous scale. To study internal consistency of the raters further, the distribution of the differences between the self-defined time windows across the 3AFC and 4AFC conditions are examined under four categories. The first category was the differences of 0 to 0.5 minute, indicating that the rater defined nearly identical time windows for the same two instances of the Process Overview queries. The second category was differences of greater than 0.5 up to 1.5 minutes, indicating that the rater defined practically the same windows with slight differences most likely due to noise. The third category was differences of greater than 1.5 up to 2.5 minutes, indicating that the rater might have defined the same or different time windows (i.e., equivocal or grey areas). The last category was difference of greater than 2.5 minutes, indicating that the raters had defined different time windows. Based on Table 3, the raters appeared to be consistent approximately 70% of the time in defining time windows for the same two instances of the Process Overview queries. Table 3 suggests slightly higher internal consistency within raters than the ICCs on defining “recently”. Internal consistency of approximately 70% agreement may be considered adequate for defining “recently”.

[8] ICC(2,1) treats raters as a random factor; whereas, ICC(3,1) treats raters as a fixed factor. In other words, ICC(2,1) is an estimate of reliability independent of the participants/raters recruited for the experiment; whereas ICC(3,1) is an estimate of reliability the specific raters in this data set.
[9] ICC(2,1) is not presented as the internal consistency estimate is only applicable for the particular judge.

Table 3: Intra-rater difference in starting times of the time windows for the same queries (between 3AFC and 4AFC response alternative in Recently time window conditions).

	Rater A		Rater B		Rater C	
Difference between start time of the self-defined time windows between 3AFC and 4AFC conditions (Δ min)	Freq	Cumulative %	Freq	Cumulative %	Freq	Cumulative %
0-0.5 (no difference)	111	49.33	96	42.67	87	38.67
1-1.5 (slight differences that were likely due to noise)	46	69.78	61	69.78	65	67.56
1.5-2.5 (differences that might be due to noise)	14	76.00	30	83.11	19	76.00
>2.5 (significant differences that were unlikely due to noise)	54		38		54	
Total number of queries	225					

Cursory comparisons between inter-rater agreement and internal consistency estimates suggest that the raters appeared more consistent with themselves than one another in defining time windows in the Recently condition. In other words, the process experts are likely to be more consistent to themselves than to one another in defining the time windows in the Recently condition.

Reviews of the time windows and discussions with raters (i.e., process experts) suggested that defining the start time of the windows for judging parameter behaviours may be based on many considerations such as the following:

- start time at the beginning of the scenario (or scenario period) when scenario events do not directly affect the parameters
- start time at the last important event even though the scenario event does directly affect the parameters
- start time at the beginning (antecedent) of an relevant event because experts judge that knowledge about parameter behaviours from time before the event to the time of freeze is critical
- start time at the end (consequence) of an relevant event because experts judge that knowledge about parameter behaviours from the time after the event to the time of freeze is critical
- start time at “a few minutes” before the freeze because experts define “recently” based on their unique operational experience with other control room operators

In the debriefing discussions, the process experts often emphasized the importance of scenario characteristics in defining “recently”. It appeared that experts relied on their mental models of the scenario more than inspecting the trend graphs in defining “recently”.

During data collection of this study, a question from both Raters A and C, who prepared part of the experiment where the source data was drawn, exemplified their reliance on knowledge about the simulator and scenarios. Specifically, Raters A and C identified that the trend graph for one parameter appeared to be inconsistent with the scenario development. The process experts quickly resolved their own question by realizing that a unique event in the scenario prevented the sensor from accurately measuring the values of that parameter. The concern raised and ultimately resolved by Raters A and C themselves highlighted the level of expertise and judgment involved in assessing parameter behaviours even in judging parameter behaviours from graphs.

4.2.2 Intra-rater consistency

The collected data permitted three imprecise estimations of intra-rater consistency. The data collection environment and procedures were not intentionally designed to provide such estimates, potentially leading to confounding factors in the results. Nevertheless, the intra-rater consistency estimates could provide insights on future research directions.

As presented in the section above, two indicators of intra-rater consistency are derived from the comparison of time windows defined in the treatment level of Recently between the 3AFC with the 4AFC response alternative conditions. First, the intra-class correlations demonstrate moderate level of internal consistency (i.e., 0.4 < ICC(3,1) < .6). Second, based on the distributions of differences between the self-defined time windows across two experimental conditions, raters defined almost identical time windows in the 3AFC as in the 4AFC conditions for 70% of the queries (Table 3). The raters appeared reasonably consistent in defining "recently" across the two response alternative conditions.

The third indicator of rater consistency was the degree of agreement for data collected from the same raters between this study and a prior reliability study [5] in the treatment combination of 3AFC response alternative and Recently time window. In the prior reliability study, Rater A and B answered the same Process Overview queries in real time during data collection of a full-scope simulator study; whereas, Rater C answered the Process Overview queries in the same office of this experiment based on trend graphs after the full-scope simulator experiment. Despite different data collection procedures and environment between the two studies, the raters demonstrated "moderate" agreement (see Appendix C for interpretation) with themselves (Table 4). Furthermore, the differences between κ's of the raters were negligible (Figure 9).

Table 4: Intra-rater agreement between two reliability studies.

Rater	Intra-rater agreement	
	κ	σ_κ
A	.543	.049
B	.572	.048
C	.604	.050

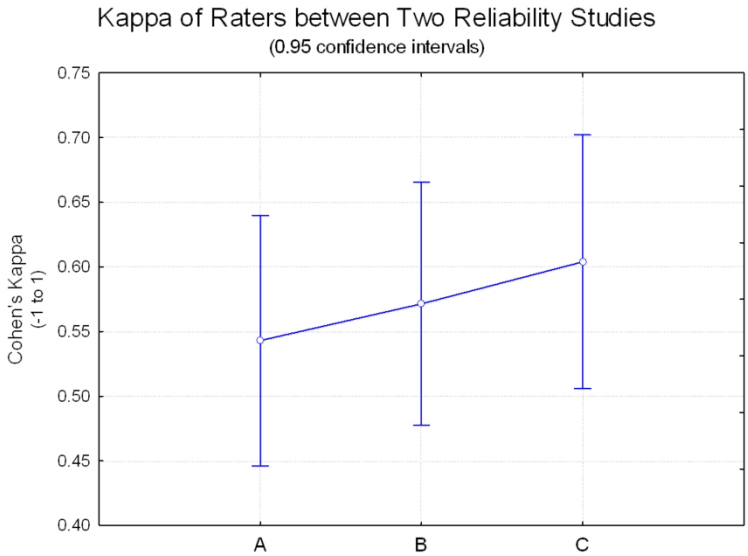


Figure 9: Intra-rater consistency across two reliability studies in Recently time windows and 3FAC response alternative treatment conditions.

In summary, the preliminary evidence indicates reasonably consistent judgment within process experts in answering Process Overview queries.

4.2.3 Preference of process experts

All three process experts indicated that the two proposed modifications to the Process Overview Measure were useful. They find that the Event-based time windows reduced ambiguities. Furthermore, they believed that recalling major events in the scenarios was acceptable when the participants would not have access to the trend graphs to respond to Process Overview queries during data collection. Their opinions on the treatment of time windows converged with the quantitative results.

For the treatment of response alternatives, the process experts believed that the extra option in the 4AFC condition could improve diagnositicity of operator monitoring performance (i.e., Process Overview). One process expert mentioned that description of the new response alternative should be revised slightly. "Fluctuating around the same point" gave the impression of parameters moving above and below a value. However, some fluctuations would be better described as varying above and below some values. Therefore, a better description could be "fluctuating near a point". Although expert opinions and the rationale for including the additional response alternative were consistent, the overall statistical results indicated that the experts implicitly disagreed on what defines fluctuations (see Figure 8).

5. DISCUSSION

5.1 Proposed modifications to the Process Overview Measure

This study intended to evaluate the effectiveness of (i) predefining time windows in the queries with events in the scenarios, and (ii) introducing a new response alternative to include parameter fluctuations to improve inter-rater agreement of the Process Overview Measure.

The results indicate that predefined time windows using scenario events (i.e., Event-based/Predefined time windows) improve agreement between process experts over undefined time windows (i.e., Recently/Undefined). During debriefing, the process experts/participants also approved of the modification. This finding is consistent with the hypothesis. Therefore, Process Overview queries should include a scenario event as the start time for eliciting operator awareness of parameter behaviours during simulator freezes.

Process Overview [4, 5] provides the conceptual basis for proposing the use of scenario events as a method to specify or anchor time periods in queries. Defining time periods according to key events in scenarios is operationally and psychologically consistent with the information processing behaviour of process operators according to field observations (e.g., [12]). Though employing absolute time scales (e.g., minutes) appears to provide an objective solution, operators often psychologically experience time from an operational perspective. In summary, the observed improvement in inter-rater agreement using scenario events to define time windows is justified both conceptually and empirically.

The exploratory analysis further illuminates the main results of this study and Process Overview (the concept). The process experts tend to define "recently" or time windows relatively consistently across two experimental conditions (i.e., 3AFC vs. 4AFC) but disagree with each other considerably, confirming that undefined time windows are indeed one factor challenging inter-rater agreement. The exploratory analysis further indicates that perception of parameter behaviours during monitoring involves substantial information processing and expertise. The exploratory analysis confirms that

process experts do consider many contextual factors in determining parameter behaviours from trend graphs (e.g., scenario descriptions) even situated outside the control room environment. The number of (at least seemingly[10]) legitimate considerations for defining "Recently" exemplifies such complex and implicit processing.

The results indicate that the additional response option of fluctuations (in the 4AFC conditions) hinders agreement between process experts over the original response alternative set (i.e., 3AFC). However, during debriefing, the process experts advocate the new response option as meaningful for monitoring. In brief, the quantitative results contradict the hypothesis and expert opinions.

According to debriefing discussions in this study (and personal communication prior to the study), the process experts agree on the conceptual usefulness of knowing fluctuations for various parameters. However, the empirical results suggest that the process experts appear to disagree, at least implicitly, on the criteria that define fluctuations. The lack of agreement on the defining characteristics of fluctuations may be a result of the inherent fuzziness related to the behavioural category. Fluctuations include increasing and decreasing values that indicate no practical changes on average. In other words, underlying characteristics of fluctuations may rely on all three of the other behavioural categories. The conceptual distinctness of fluctuating parameters in the mind of process experts is difficult to transfer onto a measurement scale in a manner that improves reliability and diagnosticity.

The exploratory analysis suggests that the process experts can answer Process Overview queries fairly consistently across different data collection settings (e.g., real time vs. post experiment). The extensive reliance on knowledge about the simulator/nuclear power plant and scenarios may have minimized the significance of the data collection environment. Though data were not collected with sufficient control for studying the impact of data collection environment in answering Process Overview queries, there are indications that post-experiment scoring with process experts is an acceptable alternative when real-time scoring becomes unfeasible.

5.2 Implication: Modification to the Process Overview Measure

The empirical results indicate that inter-rater agreement of the Process Overview Measure could improve by specifying an important process event in the scenario (or scenario period) as the starting time for judging parameter behaviours of Process Overview queries (i.e., the Event-based/Predefined time window treatment level). Therefore, future implementation of the Process Overview Measure should predefine time windows using scenario events in the queries (Figure 10).

Process Overview query:

Since event X (e.g., telephone call from field operators) until now, parameter [code] (e.g., condenser pressure) has:

Process Overview response alternatives:

(i) decreased (ii) stayed the same (iii) increased

Figure 10: Recommended version of query structure and response set for the Process Overview Measure.

[10] In many cases, it is not feasible to assess the correctness of subject matter experts.

Predefining time windows in queries is an additional step to the procedure of the Process Overview Measure. While developing scenarios for experiments or training sessions, process experts not only need to identify parameters but also construct or select scenario events to develop Process Overview queries. The scenario events should have two qualities. First, the events should be salient and central to the scenarios (or scenario periods) because the detection of the events should not become part of the assessment. In other words, all operators should be aware of the events. Examples include alarms of major systems, scram of the simulated NPP and telephones calls from field operators about equipment failures. Second, the events should be robust or independent to operator control actions because queries would become inapplicable in cases where operators prevent the intended events from occurring during the scenarios. When the intended events predefining the time windows do not occur, the queries and responses could lead to invalid measurements of Process Overview.

In addition the procedural changes to the developing Process Overview queries, process displays in the observation gallery and the data logging systems for the simulator should be updated to store and present the scenario events, respectively. The updated process displays could support process experts in judging parameter behaviours and the logging system could offer the possibility of post-experiment/data collection scoring by process experts.

5.3 Limitation

The method of this study aimed to facilitate the kinds of information processing and judgment similar to those necessary in answering Process Overview queries during data collection of full-scope simulator experiments. The exploratory data analysis demonstrates evidence of success in eliciting such complex information processing in this study. Thus, the effects of the experimental manipulations revealed in this study are expected to generalize to implementations of the Process Overview Measure for full scope simulator studies.

Both the method and data analysis, however, could not ensure that the findings are completely based on process experts or participants being consistently engaged in the same information processing or cognitive activities as in full-scope simulator environment. Relying on trend graphs and paper records to simulate parameter behaviours in dynamic situations for answering Process Overview queries could be challenging even for process experts. Thus, the process experts could be tempted to convert the experimental task, which should involve a mix of visual inspection of parameter trends and mental simulation of the contexts, into a graphical judgment task. The Event-based/Predefined condition appears particularly prone to the temptation of simplifying the experimental task. The exploratory data analysis does not indicate that the process experts have merely performed a graphical judgment task, but modification of the Process Overview Measure requires validation in full-scope simulator experiments.

5.4 Future work

Future work is necessary to validate the recommended modification of the Process Overview Measure in a representative operational environment. Inter-rater reliability studies that employ representative operational settings and multiple process experts involved in experiment preparation could validate the success of the modified Process Overview Measure. Furthermore, the feasibility of identifying events during scenario development that set the time windows for judging parameter behaviours is unknown until the modified Process Overview Measure is tried in a full-scope simulator experiments. Qualitative data from control room operators serving as participants in the full-scope simulator experiments should also be collected for further assessment of the modified Process Overview Measure.

Future work may investigate operator conceptualization of "fluctuation" in details. The empirical evidence suggests that the definition or concept of fluctuation is complex in operational settings. To successfully operationalize the idea of fluctuation, qualitative research appears necessary to fully understand the defining characteristics of fluctuating parameters and the operational utility of knowing such fluctuations.

6. CONCLUSION

The objective of this study was to improve an assessment tool for operator monitoring performance - the Process Overview Measure. The empirical study investigated two proposals: (i) predefining time windows in the query with scenario events, and (ii) adding the "fluctuating around the same point" as a category of parameter behaviours in the response set. The experiment indicates that predefining time-windows with scenarios events could improve agreement between process experts; whereas adding the "fluctuating around the same point" as a response alternative could hinder agreement (relatively to the version of the Process Overview Measure prior to this study). Therefore, the Process Overview Measure now requires process experts to specify events to be the starting times for judging parameter behaviours when they identify relevant process parameters for the queries during scenario development.

The improved inter-rater agreement between process experts for the new Process Overview Measure should reduce noise in Process Overview measurements, thereby providing more sensitive and reliable indicator of operator monitoring performance. Therefore, the improvement of the Process Overview Measure should contribute to the assessment of operator performance and control room support in monitoring NPPs.

7. REFERENCES

[1] K. J. Vicente, *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. Mahwah, N.J.: Lawrence Erlbaum Associates, 1999.

[2] N. Moray, "Oú sont les neiges d'antan?," in *Human Performance, Situation Awareness and Automation*, Daytona Beach, FL, 2004.

[3] J. Patrick and N. James, "A task-oriented perspective of situation awareness," in *A Cognitive Approach to Situation Awareness: Theory and Application*, S. P. Banbury and S. Tremblay, Eds., Hampshire, UK: Ashgate, 2004.

[4] G. Skraaning Jr*., et al.*, "The Ecological Interface Design Experiment (2005)," OECD Halden Reactor Project, Halden, Norway HWR-833, 2007.

[5] N. Lau*, et al.*, "Situation Awareness in Monitoring Nuclear Power Plants: The Process Overview Concept and Measure," OECD Halden Reactor Project, Halden, Norway HWR-954, 2010.

[6] D. N. Hogg*, et al.*, "Development of a situation awareness measure to evaluate advanced alarm systems in nuclear power plant control rooms," *Ergonomics,* vol. 38, pp. 2394-2413, 1995.

[7] D. N. Hogg*, et al.*, "Measurement of the Operator's Situation Awareness for Use Within Process Control Research: Four Methodological Studies," OECD Halden Reactor Project, Halden, Norway HWR-377, 1994.

[8] M. R. Endsley, "Direct Measurement of Situation Awareness: Validity and Use of SAGAT," in *Situation awareness: analysis and measurement*, M. R. Endsley and D. J. Garland, Eds., Mahwah, NJ: Lawrence Erlbaum Associates, 2000, pp. 147-174.

[9] M. R. Endsley, "Measurement of situation awareness in dynamic systems," *Human Factors,* vol. 37, pp. 65-84, 1995.

[10] G. Skraaning Jr., "Experimental Control versus Realism: Methodological Solutions for Simulator Studies in Complex Operating Environments.," OECD Halden Reactor Project, Halden, Norway HPR-361, 2003.

[11] J. Sims and C. C. Wright, "The Kappa Statistics in Reliability Studies: Use, Interpretation, and Sample Size Requirements," *Physical Therapy,* vol. 85, pp. 257-268, 2005.

[12] V. de Keyser, "Structuring of Knowledge of Operators in Continuous Processes: Case Study of a Continuous Casting Plant Start-up," in *New Technology and Human Error*, J. Rasmussen*, et al.*, Eds., Chichester, UK: John Wiley & Sons Ltd., 1987, pp. 247-259.

[13] M. H. R. Eitrheim*, et al.*, "Staffing Strategies in Highly Automated Future Plants Results from the 2009 HAMMLAB Experiment," OECD Halden Reactor Project, Halden, Norway HWR-938, 2010.

[14] G. Skraaning Jr*, et al.*, "Coping with Automation in Future Plants: Results from the 2009 HAMMLAB Experiment," OECD Halden Reactor Project, Halden, Norway HWR-937, 2010.

[15] A. Drøivoldsmo, "New tools and technology for the study of human performance in simulator experiments," PhD dissertation, Norwegian University of Science and Technology, Trondheim, Norway, 2003.

[16] J. Cohen, "Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," *Psychological Bulletin,* vol. 70, p. 4, 1968.

[17] J. Cohen, "The coefficient of agreement for nominal scales," *Educational and Psychological Measurement,* vol. 20, pp. 37-46, 1960.

[18] R. E. Kirk, *Experimental design: procedures for the behavioral sciences*, 3rd ed. Pacific Grove, CA: Brooks/Cole, 1995.

[19] Statsoft. (2010, May). *ANOVA/MANOVA*. Available: http://www.statsoft.com/textbook/anova-manova/#homogeneity

[20] D. Howell, *Statistical Methods for Psychology*, 5th ed. Pacific Grove, CA: Duxbury/Thomson Learning, 2002.

[21] A. Field and G. Hole, *How to design and report experiments*. Thousand Oaks, CA. USA: Sage publications Ltd, 2003.

[22] G. W. Corder and D. I. Foreman, *Nonparametric statistics for non-statisticians: a step-by-step approach*. Hokoken, NJ, USA: John Wiley & Sons, Inc., 2009.

[23] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bulletin,* vol. 86, pp. 420-428, 1979.

[24] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics,* vol. 33, pp. 159-174, 1977.

[25] P. E. Shrout, "Measurement reliability and agreement in psychiatry," *Statistical Methods in Medical Research,* vol. 7, pp. 301-317, 1998.

[26] E. Jeannot, "Situation Awareness, Synthesis of Literature Research," EUROCONTROL, Brussels, Belgium ECC Note No. 16/00, 2000.

[27] E. Jeannot*, et al.*, "The Development of Situation Awareness Measures in ATM Systems," Eurocontrol, Brussels, Belgium HRS/HSP-005-REP-01, 2003.

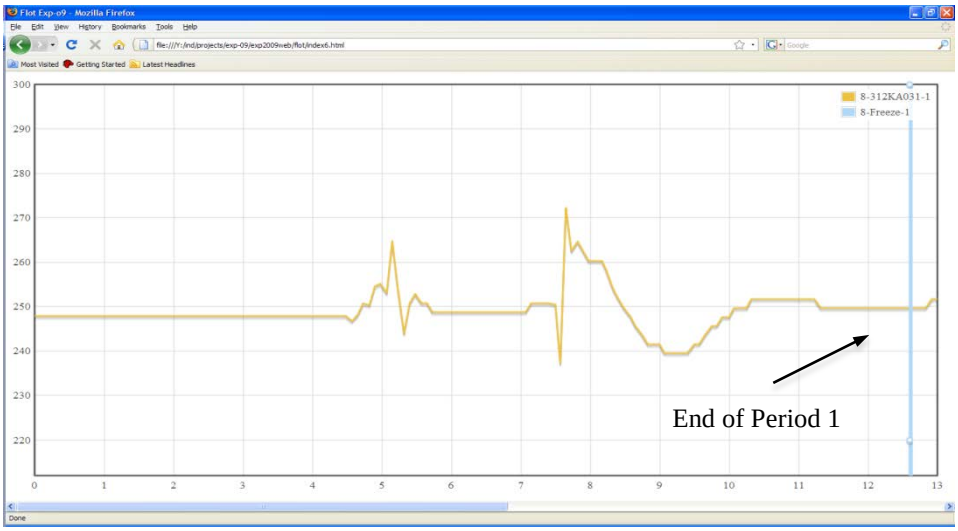
[28] R. Breton and R. Rousseau, "Situation Awareness: A review of the Concept and its Measurement," Defence Research and Development Canada, Valcartier, QC, Canada TR 2001-220, 2001.

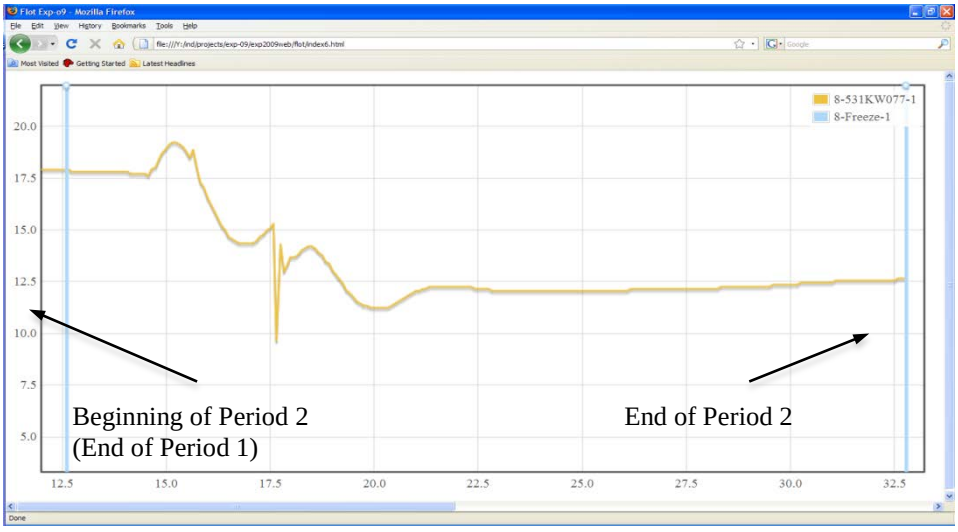
APPENDIX A: INSTRUCTION

Introduction

The aim of this study is to assess inter-rater reliability of different modifications to the Process Overview Measure. The experimental task requires inspecting graphs of various parameters and responding to multiple-choice questions about the changes of those parameters.

A typical graph looks like the following:





You will see one line for the first scenario period and two lines for the second scenario period at the ends of each graph. These lines define the scenario periods.

The following two types of questions are administered:

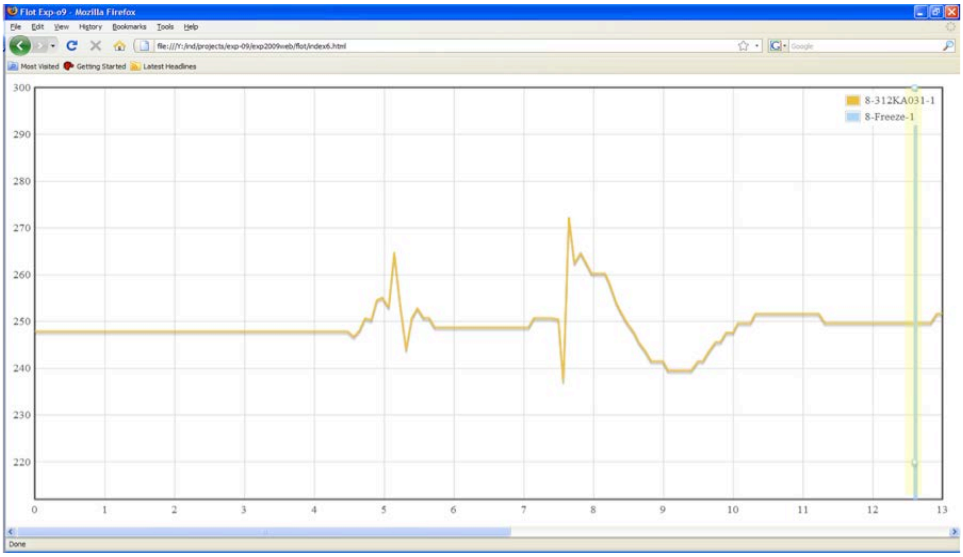
Type 1:

I den senaste tiden har summa drivdonsprocent 221KW700 (i %):

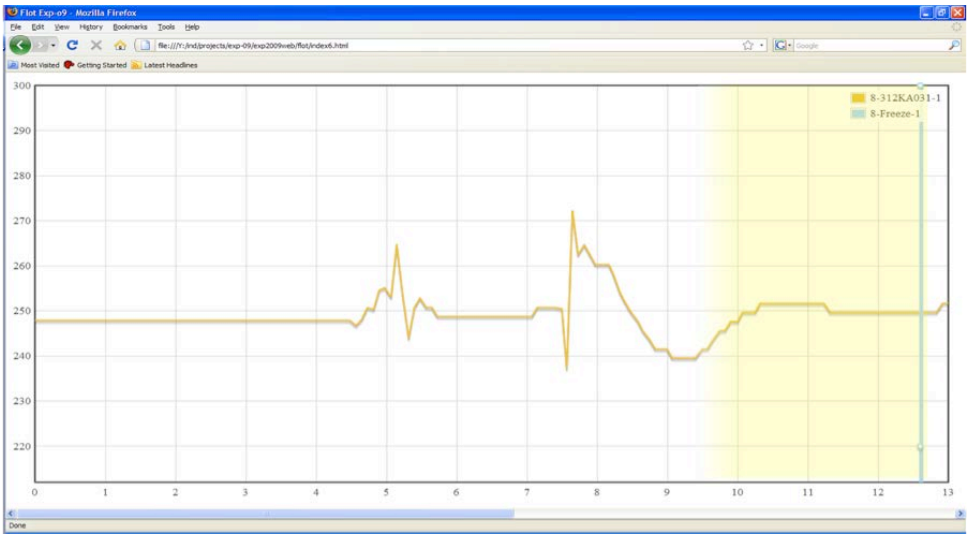
Type 2:

Från sista gången den automatiska agenten stoppade i den här perioden tills nu, har HC-flödet 211KW032:

When the question begins with “I den senaste tiden”, the graph looks like the following:



You should first click on the yellow bar/box and increase the width of the box (by dragging) to the point that covers the portion of the graph that you find **meaningful** to make your judgment about the change for that parameter (like below).

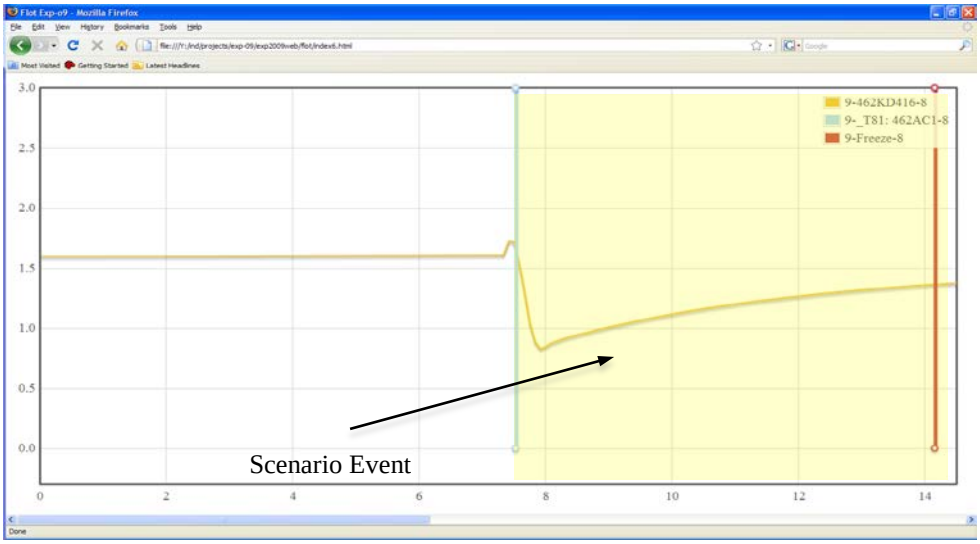


There are many factors in defining a “recent” time period that are meaningful to answer the question. The following is a non-exhaustive list of considerations:

- Scenario characteristics/descriptions
  - Process faults
  - Process events
- Operator actions (or lack thereof) according to OPAS
- Automation actions
- Alarms
- Parameters characteristics

Avoid defining the “recent” period solely by visual inspection of the graphs. It is important that the time period concerning “recently” is operationally meaningful to you.

When the question begins with a scenario event (e.g., “Från sista gången den automatiska agenten stoppade i den här perioden tills nu”), you should see an extra line on the graph that signifies the event specified in the question. For these questions, make the judgment about the parameter change between the time of the event to the end of the scenario period (as depicted by the two lines and highlighted yellow).



There are two sets of multiple-choice responses to the questions (which should be self-explanatory).

(1) minskat (2) varit konstant (3) ökat

(1) minskat (2) varit konstant (3) varierat kring samma punkt (4) ökat

Like defining recently, there are many factors to detecting **meaningful parameter changes**. In making your judgment, the following is a non-exhaustive list of considerations:

- Scenario characteristics/descriptions
  - Process faults
  - Process events
- Operator actions (or lack thereof) according to OPAS
- Automation actions
- Alarms
- Parameters characteristics (e.g., typical noise for the sensor)

Be aware of the scale (y-axis). If necessary, ask for the experimenter to rescale the graph for you.

Avoid judging the changes sole by visual inspection of the graph. It is important that the parameter change is operationally meaningful to know (rather than visually detectable). Note that this applies to the differences between the choice of ”varit knostant” and “varierat kring samma punkt”. If the fluctuation is very small without any operational significance or value, it is more meaningful to respond “varit konstant”. However, if the knowledge about the fluctuation is valuable for the operation, then “varierat kring samma punkt” is more meaningful and appropriate.

APPENDIX B: DEBRIEFING GUIDE

Agenda

1. Thank participants
2. Semi-structured interview
3. Ask participants for any other feedback and questions.
4. Inform participants to contact experimenter/HRP for any further requests/concerns.

Time Windows

Which type of query you rather have: with or without events?

Why?

Extra questions if necessary:

What do you think of the event probe in specifying a time window?

Does it help your definition of the queries or hinder your flexibility in judgment?

Do you have some examples where the event probe help and hinder you?

How do you think of time while operating nuclear power plants?

Response Alternatives

Which type of response options do you rather have: the 3AFC or 4AFC?

Why?

Extra questions if necessary:

What do you think of the additional response option for judging parameter changes?

Do you have cases where the additional or the lack of the option makes it difficult or confusing to answer the query?

Is the response set complete for operating nuclear power plant?

APPENDIX C: DESCRIPTION OF COHEN'S KAPPA

The Cohen's Kappa, κ, [11, 16, 17] is applied to assess the agreement between raters on a nominal scale. Equation 1 and 2 express κ and its standard error, σ_κ, respectively.

$$\kappa = \frac{p_o - p_c}{1 - p_c} = \frac{f_o - f_c}{N - f_c}, \quad [1]$$

$$\sigma_\kappa = \sqrt{\frac{p_o(1-p_o)}{N(1-p_c)^2}} = \sqrt{\frac{f_o(N-f_o)}{N(N-f_o)^2}}, \quad [2]$$

where (for both Equation 1 and 2) p_o=proportion of units agreed; f_o= frequency of units agreed; p_c=proportion of units agreed expected by chance; f_c=frequency of units agreed expected by chance; N denotes total sample size.

κ ranges from -1 to 1, for which the lower bound expresses complete disagreement and upper bound express complete agreement. The statistic is similar to basic correlation statistics except that κ is adjusted for the agreement by chance.

Cohen also provides Equation 3 for significance testing between two κ's :

$$z = \frac{\kappa_1 - \kappa_2}{\sqrt{\sigma_{\kappa_1}^2 + \sigma_{\kappa_2}^2}} \quad \textbf{[3]}$$

Table 5 presents the commonly accepted interpretation of κ [11] (also see e.g., [24, 25]). For tasks in complex environments involving substantial judgment (e.g., medical diagnosis), κ's above 0.4 is acceptable and generally do not exceed 0.7.

Table 5: Commonly accepted interpretation of Cohen's Kappa.

Kappa coefficient	Strength of agreement
$\kappa \leq 0$	Poor
$0 < \kappa \leq .2$	Slight
$.2 < \kappa \leq .4$	Fair
$.4 < \kappa \leq .6$	Moderate
$.6 < \kappa \leq .8$	Substantial
$.8 < \kappa \leq 1$	Almost perfect

APPENDIX D: KAPPA ANALYSIS WITH ALTERNATIVE DATA AGGREGATION PROCEDURE

Table 6 and Figure 11 summarize the results in terms of κ calculated per experimental conditions[11] (see Appendix C for a formula for calculating κ statistics).

Table 6: Cohen's Kappa calculated per experimental conditions.

	AB		AC		BC	
	κ	σ_κ	κ	σ_κ	κ	σ_κ
Recently & 3AFC	0.497	0.050	0.467	0.052	0.308	0.056
Recently & 4AFC	0.404	0.046	0.429	0.045	0.231	0.048
Event & 3AFC	0.780	0.035	0.541	0.048	0.424	0.048
Event & 4AFC	0.610	0.041	0.431	0.045	0.408	0.044

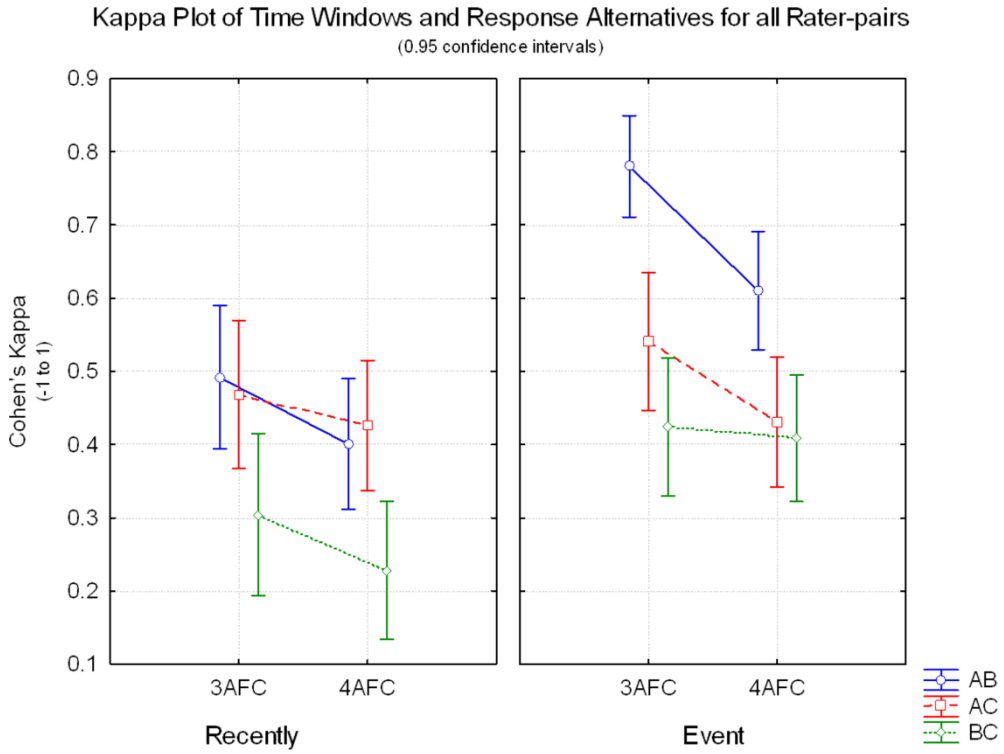


Figure 11: Plot of kappa calculated per experimental conditions.

To varying degrees, the rater-pairs demonstrated similar treatment effects - higher agreement with Event-based time windows and 3AFC response alternatives than Recently time windows and 4AFC response alternatives, respectively. Given the general consistency between raters, κ's were calculated neglecting rater-pairs and interactions to examine main effects only.

[11] Note that the data aggregation and confidence intervals for Cohen's κ have different properties from some other scales (e.g., response time). The data aggregation and variance calculation follow Cohen's formula in Appendix C as opposed to the standard mean and variance formula for ANOVAs. The data aggregation and variance calculation are intended for pair-wise comparisons using the Z-test (see Appendix C). For this reason, the means and confidence intervals are plotted for statistical inferences, even though the data set is treated as repeated-measures/RBF design. Typically, graphical plots of means and confidence intervals for measurements other than κ does not lead to correct statistical inferences for repeated-measures design because the confidence intervals does not appropriately account for within-subject variance.

Following the Holm's family-wise error-rate adjustment for multiple comparisons [20], Cohen's significance testing procedure (i.e., Equation 3 in Appendix C) indicates both main effects to be significaist. Specifically, the Event-based/Predefined (κ=0.553, σ_κ=0.017) lead to significantly higher agreement between raters than Recently/Undefined (κ=0.423, σ_κ=0.019) time window (Z=5.00, p=.00); and the 4AFC (κ=0.421, σ_κ=0.019) lead to significantly lower agreement between raters than 3AFC (κ=0.507, σ_κ=0.020) response alternative (Z=3.16, p=.00). Figure 12 and Figure 13 illustrates the main effects of time windows and response alternatives, respectively.

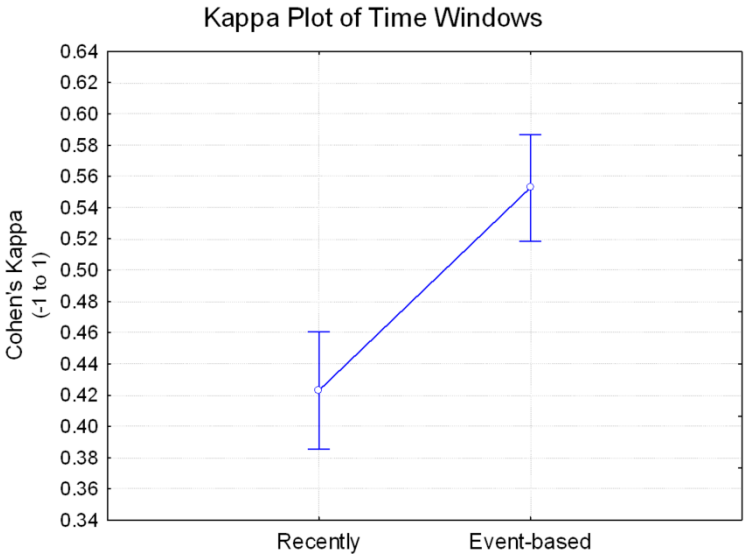


Figure 12: Plot of aggregated Cohen's Kappa for the main effect of Time Window.

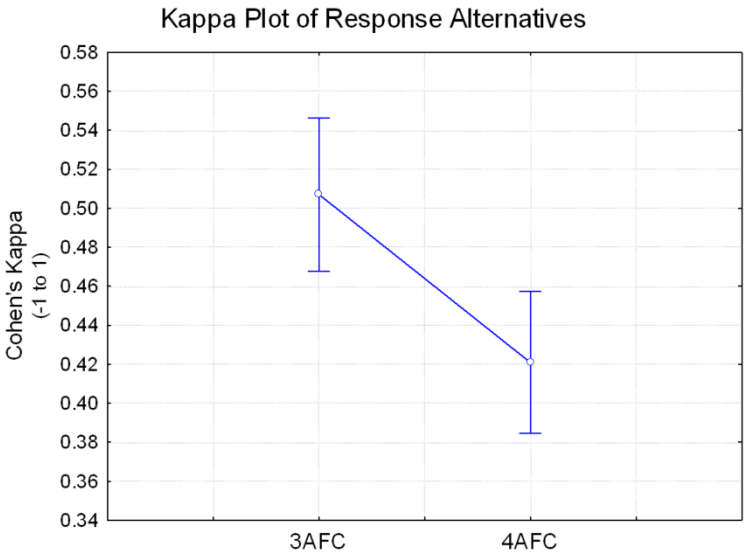
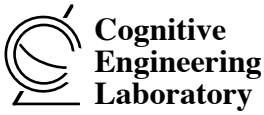


Figure 13: Plot of aggregated Cohen's Kappa for the main effect of Response Alternative.

In summary, the analysis with κ's led to two findings: (i) the Event-based improves reliability over Recently time windows (consistent with hypothesis), and (ii) 4AFC hinder reliability compared to 3AFC response alternatives (in contradiction with hypothesis).



CEL TECHNICAL REPORT SERIES

CEL 93-01 "Egg-sucking, Mousetraps, and the Tower of Babel: Making Human Factors Guidance More Accessible to Designers"
• Kim J. Vicente, Catherine M. Burns, & William S. Pawlak

CEL 93-02 "Effects of Expertise on Reasoning Trajectories in an Abstraction Hierarchy: Fault Diagnosis in a Process Control System"
• Klaus Christoffersen, Alex Pereklita, & Kim J. Vicente

CEL 94-01 "Cognitive 'Dipsticks': Knowledge Elicitation Techniques for Cognitive Engineering Research"
• Klaus Christoffersen, Christopher N. Hunter, & Kim J. Vicente

CEL 94-02 "Muddling Through Wicked Problems: Exploring the Role of Human Factors Information in Design"
• Catherine M. Burns

CEL 94-03 "Cognitive Work Analysis for the DURESS II System"
• Kim J. Vicente & William S. Pawlak

CEL 94-04 "Inducing Effective Control Strategies Through Ecological Interface Design"
• William S. Pawlak

CEL 94-05 "Research on Factors Influencing Human Cognitive Behaviour (I)"
• Klaus Christoffersen, Christopher N. Hunter, & Kim J. Vicente

CEL 94-06 "Ecological Interfaces for Complex Industrial Plants"
• Nick Dinadis & Kim J. Vicente

CEL 94-07 "Evaluation of a Display Design Space: Transparent Layered User Interfaces"
• Beverly L. Harrison, Hiroshi Ishii, Kim J. Vicente, & Bill Buxton

CEL 94-08 "Designing and Evaluating Semi-Transparent 'Silk' User Interface Objects: Supporting Focused and Divided Attention"
• Beverly L. Harrison, Shumin Zhai, Kim J. Vicente, & Bill Buxton

CEL 95-01 "An Ecological Theory of Expertise Effects in Memory Recall"
• Kim J. Vicente & JoAnne H. Wang

CEL-SP "Strategic Plan"
• Cognitive Engineering Laboratory

CEL-LP "Cognitive Engineering Laboratory Profile"
• Cognitive Engineering Laboratory

CEL 95-04 "A Field Study of Operator Cognitive Monitoring at Pickering Nuclear Generating Station-B"
• Kim J. Vicente & Catherine M. Burns

CEL 95-05 "An Empirical Investigation of the Effects of Training and Interface Design on the Control of Complex Systems"
• Christopher N. Hunter

CEL 95-06 "Applying Human Factors to the Design of Medical Equipment: Patient-Controlled Analgesia"
• Laura Lin, Racquel Isla, Karine Doniz, Heather Harkness, Kim J. Vicente, & D. John Doyle

CEL 95-07 "An Experimental Evaluation of Transparent Menu Usage"
• Beverly L. Harrison & Kim J. Vicente

CEL 95-08 "Research on Factors Influencing Human Cognitive Behaviour (II)"
• Christopher N. Hunter, Michael E. Janzen, & Kim J. Vicente

CEL 95-09 "To the Beat of a Different Drummer: The Role of Individual Differences in Ecological Interface Design"
• Dianne Howie

CEL 95-10 "Emergent Features and Temporal Information: Shall the Twain Ever Meet?"
• JoAnne H. Wang

CEL 95-11 "Physical and Functional Displays in Process Supervision and Control"
• Catherine M. Burns & Kim J. Vicente

CEL 96-01 "Shaping Expertise Through Ecological Interface Design: Strategies, Metacognition, and Individual Differences"
• Dianne E. Howie

CEL 96-02 "Skill, Participation, and Competence: Implications of Ecological Interface Design for Working Life"
• Peter Benda, Giuseppe Cioffi, & Kim J. Vicente

CEL 96-03 "Practical Problem Solving in a Design Microworld: An Exploratory Study'
• Klaus Christoffersen

CEL 96-04 "Review of Alarm Systems for Nuclear Power Plants"
• Kim J. Vicente

CEL 96-05 "DURESS II User's Manual: A Thermal-hydraulic Process Simulator for Research and Teaching"
• Lisa C. Orchanian, Thomas P. Smahel, Dianne E. Howie, & Kim J. Vicente

CEL 96-06 "Research on Factors Influencing Human Cognitive Behaviour (III)"
• Dianne E. Howie, Michael E. Janzen, & Kim J. Vicente

CEL 96-07 "Application of Ecological Interface Design to Aviation"
• Nick Dinadis & Kim J. Vicente

CEL 96-08 "Distributed Cognition Demands a Second Metaphor for Cognitive Science"
• Kim J. Vicente

CEL 96-09 "An Experimental Evaluation of Functional Displays in Process Supervision and Control"
• Catherine M. Burns and Kim J. Vicente

CEL 96-10 "The Design and Evaluation of Transparent User Interfaces: From Theory to Practice"
• Beverly L. Harrison

CEL 97-01 "Cognitive Functioning of Control Room Operators: Final Phase"
• Kim J. Vicente, Randall J. Mumaw, & Emilie M. Roth

CEL 97-02 "Applying Human Factors Engineering to Medical Device Design: An Empirical Evaluation of Two Patient-Controlled Analgesia Machine Interfaces"
• Laura Lin

CEL 97-03 "ADAPT User's Manual: A Data Analysis Tool for Human Performance Evaluation in Dynamic Systems"
• Xinyao Yu, Farzad S. Khan, Elfreda Lau, Kim J. Vicente, & Michael W. Carter

CEL 97-04 "Research on the Characteristics of Long-Term Adaptation"
• Xinyao Yu, Renée Chow, Greg A. Jamieson, Rasha Khayat, Elfreda Lau, Gerard Torenvliet, Kim J. Vicente, & Michael W. Carter

CEL 97-05 "A Comprehensive Experimental Evaluation of Functional Displays in Process Supervision and Control"
• Catherine M. Burns and Kim J. Vicente

CEL 98-01 "Applying Human Factors Engineering to Medical Device Design: An Empirical Evaluation of Patient-Controlled Analgesia Machine Interfaces"
• Laura Lin

CEL 98-02 "Building an Ecological Foundation for Experimental Psychology: Beyond the Lens Model and Direct Perception"
• Kim J. Vicente

CEL 98-03 "Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-based Work"
• Kim J. Vicente

CEL 98-04 "Ecological Interface Design for Petrochemical Processing Applications"
• Greg. A. Jamieson & Kim J. Vicente

CEL 98-05 "The Effects of Spatial and Temporal Proximity of Means-end Information in Ecological Display Design for an Industrial Simulation"
• Catherine M. Burns

CEL 98-06 "Research on Characteristics of Long-term Adaptation (II)"
• Xinyao Yu, Gerard L. Torenvliet, & Kim J. Vicente

CEL 98-07 "Integrated Abstraction Hierarchy and Plan-Goal Graph Model for the DURESS II System: A Test Case for Unified System- and Task-based Modeling and Interface Design"
• Christopher A. Miller & Kim J. Vicente

CEL 98-08 "Comparative Analysis of Display Requirements Generated via Task-Based and Work Domain-based Analyses: A Test Case Using DURESS II"
• Christopher A. Miller & Kim J. Vicente

CEL 98-09 "Abstraction Decomposition Space Analysis for NOVA's E1 Acetylene Hydrogenation Reactor"
• Christopher A. Miller & Kim J. Vicente

CEL 99-01 "Development of an Analytical Framework and Measurement Tools to Assess Adaptive Performance of Individual and Teams in Military Work Domains"
• John R. Hajdukiewicz, Kim J. Vicente, & Robert G. Eggleston

CEL 99-02 "Applying Perceptual Control Theory and Ecological Interface Design to the Control Display Unit"
• Sandra Chéry

CEL 99-03 "Research on the Characteristics of Long-Term Adaptation (III)"
• Gerard L. Torenvliet & Kim J. Vicente

CEL 99-04 "Comparative Analysis of Display Requirements Generated via Task-Based and Work Domain-based Analyses in a Real World Domain: NOVA's Acetylene Hydrogenation Reactor"

• Christopher A. Miller & Kim J. Vicente

CEL 99-05 "A Cognitive Engineering Approach for Measuring Adaptive Behavior"
• John R. Hajdukiewicz & Kim J. Vicente

CEL 00-01 "Differences Between the Eye-fixation Patterns of Novice and Expert Operators of the DURESS II Physical Interface"
• Madhava Enros & Kim J. Vicente

CEL 00-02 "If Technology Doesn't Work for People, then It Doesn't Work"
• Kim J. Vicente

CEL 01-01 "A Field Study of Collaborative Work in Network Management: Implications for Interface Design and Evaluation"
• Renée Chow & Kim J. Vicente

CEL 01-02 "A Prototype Ecological Interface for a Simulated Petrochemical Process"
• Greg A. Jamieson & Wayne H. Ho

CEL 01-03 "EID Design Rationale Project: Case Study Report"
• Greg A. Jamieson, Dal Vernon C. Reising & John Hajdukiewicz

CEL 02-01 "Ecological Interface Design for Petrochemical Process Control: Integrating Task-and System-Based Approaches"
• Greg A. Jamieson

CEL 06-01 "Canada Foundation for Innovation (CFI)
Emerson DeltaV / MiMiC
Industrial Process Control Simulator"
• Antony Hilliard & Laura Thompson

CEL 06-02 "Developing Human-Machine Interfaces to Support Monitoring of UAV Automation"
• Greg A. Jamieson, Lu Wang, Jamy Li

CEL 06-03 "Sensor Noise and Ecological Interface Design: Effects of Noise Magnitude on Operators' Performance and Control Strategies"
• Olivier St-Cyr

CEL 07-01 "The 2005 Ecological Interface Design Process and the Resulting Displays"
• Robin Welch, Alf Ove Braseth, Christer Nihlwing, Gyrd Skraaning, Arild Teigen, Øystein Veland, Nathan Lau, Greg A. Jamieson, Jordana Kwok, Catherine M. Burns

CEL 07-02 "The Ecological Interface Design experiment (2005)"
• Gyrd Skraaning, Nathan Lau, Robin Welch, Christer Nihlwing, Gisle Andresen, Liv Hanne Brevig, Øystein Veland, Greg A. Jamieson, Catherine M. Burns, Jordanna Kwok

CEL 07-03 "Factors Influencing the Reliance on Combat Identification Systems"
• Lu Wang

CEL 07-04 "Applying a Formative Ecological Framework to Simulator Design Challenges"
• Antony Hilliard

CEL 08-01 "Cognitive Work Analysis of the City of Toronto Winter Maintenance Program"
• Laura Thompson, Antony Hilliard, & Cam Ngo

CEL 08-02 "The Impact of Ecological Displays on Operator Task Performance and Workload"
• Nathan Lau, Gyrd Skraaning jr., Greg A. Jamieson, & Catherine M. Burns

CEL 11-01 "Situation Awareness in Monitoring Nuclear Power Plants – The Process Overview Concept and Measure"
• Nathan Lau, Gyrd Skraaning jr., Maren H. R. Eitrheim, Tommy Karlsson, Christer Nihlwing, & Greg A. Jamieson

CEL 11-02 "The Process Overview Measure: Methodological Developments to Enhance Inter-Rater Reliability"
• Nathan Lau, Gyrd Skraaning jr., Maren H. R. Eitrheim, Tommy Karlsson, Christer Nihlwing, & Greg A. Jamieson