INTER-RATER RELIABILITY OF EXPERT-BASED PERFORMANCE MEASURES

Nathan Lau
Department of Systems and Information Engineering
University of Virginia
PO Box 400747, Charlottesville, VA, USA 22904
nathan.lau@virginia.edu

Greg A. Jamieson
Department of Mechanical and Industrial Engineering
University of Toronto
5 King's College Road, Toronto, ON, Canada M5S 3G8
jamieson@mie.utoronto.ca

Gyrd Skraaning Jr.
Industrial Psychology Division
OECD Halden Reactor Project
P.O. Box 173, Halden, Norway NO-1751
gyrd.skraaning@hrp.no

ABSTRACT

Human performance measures used in representative settings often rely heavily on process experts who possess the domain knowledge to evaluate operator actions and system states. This reliance on process experts may lead to inter-rater reliability measurement concerns. We report an empirical study that assessed the inter-rater reliability of two expert-based performance measures employed in a full scope nuclear power plant simulator experiment. The results indicated low inter-rater reliability for measures employing generic cognitive performance anchors that were designed to account for dynamics introduced by operators. In contrast, results indicated high inter-rater reliability for measures employing on pre-defined, specific behavioral anchors that were inflexible to unintended changes to the progression of the scenario. Future research is necessary to develop human performance measures that are both reliable across experts and flexible to dynamics introduced by operators.

Key Words: Human Performance measures, inter-rater reliability

1 INTRODUCTION

Human performance measures used in representative settings often rely heavily on process experts. Researchers and practitioners typically (i) do not have the technical and operational expertise to assess operator control actions in complex systems, and (ii) cannot fully predict operator responses to scenario events and the resulting system behaviors for which the performance indicators should account. Thus, process experts are often present in representative experiments and certification tests to provide real-time human performance assessment. Process experts might determine the operating plant (simulator) states, describe operator rationales and responses to scenario events for qualitative assessment, and/or rate operators on performance scales. However, process experts could introduce subjectivity in human performance measurements leading to poor inter-rater reliability (i.e., disagreement between experts) and

unknown bias in the assessment (e.g., [1]). Inter-rater reliability is therefore an important indicator to support interpretation of measurements that rely on expert judgment.

This paper presents an elementary study that evaluated the inter-rater reliability of two human performance measures used in a full-scope nuclear power plant simulator experiment. The objective of this study is two-fold. First, we need to assess the inter-rater reliability of the two measures for determining adequacy for future applications. Second, we want to study whether the two measurement methods differ inter-rater reliability for informing development of human performance measures.

2 METHOD

The study was conducted at the OECD Halden Reactor Project (HRP), an international consortium for advancing technology in nuclear power generation. HRP conducts full-scope simulator experiments to evaluate design solutions and study human performance. The objective of the research program is to develop and evaluate new and modernized control rooms. In one full-scope simulator experiment, we collected inter-rater reliability data by recruiting multiple process experts to conduct human performance ratings.

This section describes in five parts the method of collecting the inter-reliability data. The first part describes the raters – the process experts – whose qualifications are particularly relevant to interpreting information about the targets of the rating task. The second part describes the rating instruments – the situation awareness and task performance measures – that were designed to capitalize on expert ratings. The third part describes the rating targets – observable operator behaviors in the experimental trials – that require domain expertise to interpret due to the complexity of representative environment. The fourth and fifth parts describe the environment and procedure for the experts to perform the rating task, respectively.

2.1 Raters/Process Experts

Two process experts (PEs) participated in this study. PE A was recruited to support full-scope simulator research activities (for two years leading up to the full-scope simulator experiment). He worked as a control room operator in multiple nuclear plants for fifteen years and participated as a process expert to develop full-scope simulator used in this experiment (see Section 2.3). He designed all of the scenarios for this experiment and advised other scientists on the processes and operations of the physical and simulator plant. He also supported the development of the situation awareness measure (see Section 2.2.1) and implementation of the task performance measure (see Section 2.2.2). During data collection, PE 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 performance measures.

PE B was a recently retired (within one year of the study) shift supervisor at the plant that the full-scope simulator replicates. He worked as a control room shift supervisor for over three decades and participated in about ten HRP studies prior to this study. He was a participant in the pilot trials of the full-scope simulator experiment.

Both PEs demonstrated substantial experience working in nuclear power plants as control room operators. Therefore, they were qualified to be the process experts assessing human performance in representative nuclear process control settings. However, PE B had less experience with the full-scope simulator and details of experiment than PE A. These differences may be relevant to the interpretation of results.

2.2 Rating Instruments

Two expert-based, human performance measures were evaluated in this study.

2.2.1 Situation Awareness – Scenario Understanding

The first measure was the Automation and Scenario Understanding Rating Scales (ASURS), which assessed operator Scenario Understanding [2], or the situation awareness specifically with regards to understanding of automation in relation to scenario events.

ASURS was composed of four scales, each with distinct, generic anchors pre-defined by researchers to guide process expert(s) on rating operator understanding of automation characteristics, scenario events and their interactions. The first scale assessed the operator awareness of automation purpose (for the scenario) on a four-point anchored scale. The second and third scales assessed operator awareness of process (or actions) and performance of automation on a five-point anchored scale. The final scale assessed operator understanding of the overall situation on a five-point anchored scale.

During the experiment, a process expert familiar with the scenario events and simulator performed ratings based on observation of operator verbalization, physical behaviors, problem solving, and system from the experimental gallery. The process expert completed the ratings once per scenario period and the final score was the sum of all the ratings. The final ASURS score reflected the level of operator understanding of the automation and events demonstrated over the scenario period as holistically judged by a process expert.

2.2.2 Task Performance

The second measure is the Operator Performance Assessment System (OPAS), which assesses operator task performance [3-5] or the effectiveness of control room operations demonstrated by the operators. OPAS provides a structure for the assessment of whether operators carry out their task work in accordance with scenario solutions prescribed a priori by experts in control room operation.

Prior to data collection, process experts analyzed the scenarios and developed optimal solutions by identifying items that expressed the desired performance. These items could differentiate between levels of task performance across experimental conditions relating to omissions, commissions, response time, strategies, and/or other performance categories. (For this study, the OPAS items were separated into *detection* and *operation* categories to isolate specific aspects of task performance.) A simple scoring system is employed, where the operators earn points for completing performance items. Each item depicts alternative operator activities that are rewarded by 0, 1, 2 or 3 points.

During the experiment, a process expert registered/rated the points earned by operators in completing the predefined activities within each performance item based on observations of operator verbalization, physical behaviors, problem solving, and system states. The final task performance score was the sum of points earned by the operators divided by the maximum available points in each scenario period with a scale ranging from 0-1. The OPAS score reflected the degree of conformance between operator performance and predefined optimal solutions to scenarios.

2.3 Rating Targets

The process experts rated performance in a full-scope simulator experiment, which recruited nine crews of licensed operators. For ASURS, the targets were all the crew behaviours observable in the control room that enabled the process experts to infer holistically the operators' level of understanding of the scenario events and automation characteristics according to the anchors in ASURS. For OPAS, the targets were overt operator activities that accomplished the specific monitoring and control actions predefined to be necessary to resolve process faults in the scenarios. The experiment employed a full scope boiling water reactor simulator [6], simulating the power generation process of a 1200MW nuclear power plant. The crews, each composed of three operators with different roles, were to operate the nuclear power plant simulator safely and efficiently for eight scenarios as if they were on duty. Each scenario introduced a unique set of process faults (e.g., a stuck valve, leaks, pump failures) at different times to challenge the crews in operating the simulated plant and interacting with their support tools (e.g.,

automatic agents). Every scenario trial was divided into two periods by a freeze for administering queries and other questionnaires (Figure 1). Because the crews did not always respond to the process faults in the same manner given the problem and solution space, the experiences of the same scenario were rarely identical across crews. (For details of the full-scope simulator experiment, see [7, 8].)

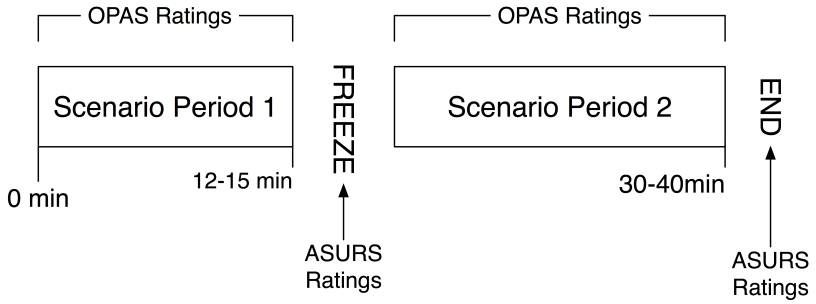

Figure 1. Structure of the scenario.

2.4 Rating Environment

The inter-rater reliability data were collected in the observation gallery (Figure 2) simultaneously with the data collection for the full-scope simulator study. Two PEs (i.e., PE A & C) were responsible for operating the simulator and plant automation. They also implemented event sequences for the scenarios. PE A acted as field operator, electrician, plant management and other roles as necessary by responding to phone calls from the control room. PE A was also the designated rater for both ASURS and OPAS as well as other performance scales. PE C was responsible for managing technical operations of the simulator associated with the experiment. An experimental leader (EL) and a laboratory technician (LT) were also present in the gallery to perform other research tasks for the experiment.

A third process expert (PE B) was recruited to provide another set of reference keys to the SA queries for the inter-rater reliability study. PE B was given a workstation between the EL and LT. The workstation (Figure 3) provided PE B with (i) two displays for process information, (ii) a pair of earbuds for audio feed from the control room, (iii) a laptop for inputting scores/ratings, (iv) a pair of binoculars for observing participants close up, (v) descriptions of all the scenarios, (vi) pen and paper for note taking, and (vii) a pair of ear muffs for muting any discussion on operator performance in the observation gallery. Furthermore, an opaque, black curtain was installed to block the raters' lines of sight. This was to reduce the likelihood of a process expert seeing (i) another process expert's monitor (used for collecting their responses to questionnaires) and (ii) any body language expressed from judging operator performance.

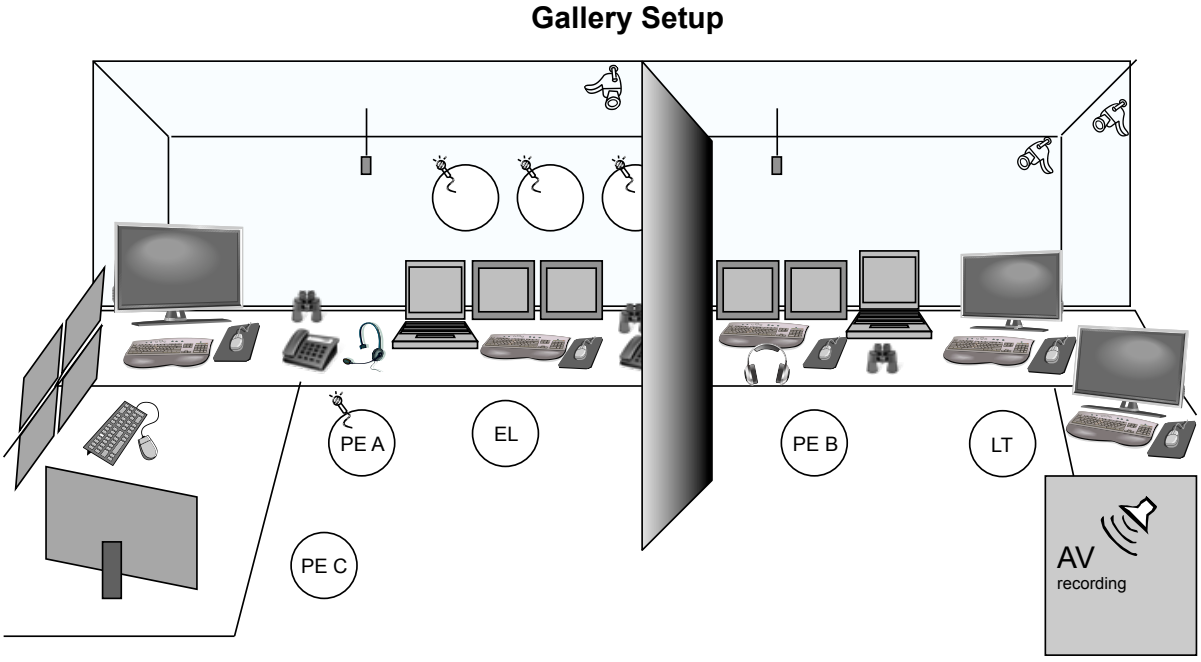

Figure 2. Observation gallery of the full-scope simulator experiment where inter-rater reliability data was collected. PE A, B & C = process expert A, B and C (PE C did not answer SA queries in Part A.); EL= experimental leader; LT = laboratory technician.

Figure 3. Workstation for PE B (the recruited shift supervisor).

2.5 Rating Procedure

PE A was given a verbal introduction about the inter-rater reliability study and instructed to withhold all discussion about all the measures and operator performance from PE B. He had received instruction on ASURS and OPAS in the preparation phase of the full-scope simulator experiment. Therefore, he was not given any further instruction, as he was the process expert responsible for performing all the human performance ratings in the full-scope simulator experiment. In other words, his responsibilities in this study were a subset of the full-scope simulator experiment.

PE B was responsible for performing human performance ratings for an additional set of data for sixteen trials of two crews in the full-scope simulator experiment. He received documents on the measures, scenarios, and the simulator for preparation two weeks prior to the study. For onsite data collection, instructions for PE B were given in two sessions. The first session consisted of a PowerPoint presentation providing an overview of the inter-rater reliability study. It also included a review of the inter-rater reliability concept and explicit instruction to refrain from discussing measurements with anyone (except the experimenter).

The second session began by reviewing the rating procedure that was to be performed concurrent with the full-scope simulator experiment. Theoretical descriptions of human performance constructs were omitted. PE B first received samples of ASURS and OPAS items, followed by a short PowerPoint presentation describing the two measures. The second session ended by familiarizing PE B with the apparatus and data collection tools in the observation gallery.

PE B was seated at his workstation and performed the ratings by entering his response with mouse and keyboard into the Halden Questionnaire System (in the same manner as PE A). PE B was new to the role of performance rater in an experimental setting; therefore the first two trials were treated as practice sessions, with scores omitted from the data analysis. In total, we collected 32 individual ASURS and OPAS performance scores and four scores for those individual measures were omitted from the analysis. After completing the performance rating of the scenario trials, the experimenter held a debriefing session with the process experts individually to elicit general comments on the measures.

3 ANALYSIS & RESULTS

We employed two types of intra-class correlation coefficients (ICC) [9] to assess the inter-rater reliability or systematic variance on an interval scale between the process experts for ASURS and OPAS. The two types are ICC(2,1) and ICC(3,1), which describe reliability of exchangeable (i.e., raters as a random factor) and non-exchangeable (i.e., raters as a fixed factor) raters, respectively. In other words, ICC(2,1) estimates *how interchangeable are the raters*; whereas, ICC(3,1) estimates *how consistent are the ratings*.

Each ICC was calculated based on 28 pairs of process expert ratings from fourteen scenario trials, each containing two scenario periods (i.e., 14 trials x 2 periods). Table I presents the ICCs for ASURS and OPAS.

Table I: Intra-class correlations of ASURS and OPAS.

	ASURS	OPAS
ICC(2,1)	.563	.825
ICC(3,1)	.585	.831

For both ASURS and OPAS, the ICC(3,1) was higher than ICC(2,1) but the difference was negligible. To account for variability of the entire rater population, ICC(2,1) is expected to be lower than ICC(3,1), which only accounts for specific raters. Given the negligible magnitude difference between the two ICC types, the main results are the ICC(2,1) statistics indicating low reliability of ASURS (ICC(2,1) = .56) and high reliability of OPAS (ICC(2,1) = 0.82). (For reference, [1] treats measures with approximately .6 as prospective measures and .9 as well empirically supported measures, such as an IQ test.) These results illustrate that expert inferences differ on generic, cognitive performance items but converge on observation of predefined, perceptible behaviors.

4 DISCUSSION

4.1 Inter-rater reliability

ASURS relied on process experts to rate operator understanding of the scenario events, automation behaviors and their interaction on a generic set of performance anchors applicable to most operating conditions of nuclear power plants. The empirical results showed low inter-rater reliability for ASURS, indicating that the performance anchors were *not* reliably guiding the interpretation of operator cognitive performance across experts. Debriefing with the process experts suggested that they had to speculate on the rationale of the operator behaviors (e.g., communication, control actions) in order to perform the ratings on operator understanding. We suspect that the poor inter-rater reliability of ASURS is a result of requiring experts to speculate about the rationale of operator behaviors as many rationales would appear applicable or logical for a given action.

OPAS relied on process experts to rate operator task performance with respect to a pre-defined set of observable operator behaviors deemed mandatory and optimum for responding to different scenario events. The empirical results showed high inter-rater reliability for OPAS, indicating that the pre-defined set of observable behaviors was reliably guiding the interpretation of operator task performance across experts. Because the pre-defined task performance anchors in OPAS were based on observable and specific actions, the process experts needed to be attentive to operator activities rather than making inferences from their observation during the experimental trials. For instance, the process experts needed to observe whether the operators open the specific valves in response a scenario event but did not need to make any substantial inference on the observation to perform a rating for the OPAS item. We believe that the high inter-rater reliability of OPAS is a result of the specificity and observability of the behavioral anchors.

Interpreting the inter-rater reliability results with respect to methodological details of ASURS and OPAS highlights the constraints of - and trade-off between - two human performance measurement methods. On the one hand, measures based on generic anchors afford flexibility in accommodating unintended scenario (or situation) changes due to unexpected operator responses to scenario events. However, such measurement instruments are prone to disagreement between experts; thus, the empirical data must either be accompanied with inter-rater reliability statistics, or be interpreted with caution. On the other hand, performance scores based on predefined, specific, observable behaviors are likely to be reliable across experts. However, such measurement instruments may misrepresent human performance on occasions when the predefined items cannot appropriately account for unanticipated scenario changes introduced by the operators; thus, such empirical data must be interpreted in consideration of potential exceptions (i.e., occasional invalid data points). In essence, generic, cognitive anchors could capture dynamics introduced by the operators at the risk of reliability, while pre-defined, behavior-specific anchors could provide consistency across experts at the risk of validity.

4.2 Limitations

This inter-rater reliability study has two key limitations. First, we could only recruit two process experts to act as raters in this study. Increasing the number of process experts would improve the

representativeness of, and our confidence in, the inter-rater reliability statistics. Nevertheless, the empirical results should be credible as ICC(2,1) describes reliability of exchangeable scorers. In other words, the statistics adjust for the variability beyond the recruited raters. Though limiting in terms of access, professional operators and process experts embed their domain expertise in their inferences on operator performance from their observation that could improves the validity of the collected data in representative settings.

Second, this study only examined two expert-based performance measures developed specifically for the process control domain. Thus, while the empirical results contrast inter-rater reliability of expert ratings on generic, cognitive performance anchors in comparison to specific, predefined ones, they cannot represent the degree of inter-rater reliability for all expert-based performance measures following similar approaches.

These two limitations essentially challenge generalization of the empirical results. However, neither limitation can be practically resolved in any single study in a representative setting. Thus, the likely solution to this paucity of the reliability data is continual empirical assessment of expert-based performance measures whenever they are employed.

5 CONCLUSION

This empirical study examined the inter-rater reliability of two expert-based human performance measures that applied different methodologies. The first measure – ASURS – relied on process experts to rate operator understanding of the scenario events, automation properties and their interaction on a generic set of performance anchors applicable to most operating conditions of nuclear power plants. The second measure – OPAS – relied on process experts to rate operator task performance with respect to a pre-defined set of observable operator behaviors deemed mandatory and optimum for responding to different scenario events. ASURS and OPAS exhibited low and high inter-rater reliability, respectively. The empirical results highlighted the constraints of and trade-off between two measurement methods. Generic performance anchors designed to account for unanticipated scenario changes introduced by operators are prone to reliability issues; whereas, pre-defined, scenario-specific performance anchors designed to minimize disagreement across experts are prone to validity issues. Further research is necessary to develop measures that are reliable across process experts while accommodating unintended scenario changes introduced by operator actions.

6 ACKNOWLEDGMENTS

This research was supported through a grant from the Natural Science and Engineering Research Council and internal funding of the OECD Halden Reactor Project. We are indebted to Maren H. Rø-Eitrheim for support the recruitment of the process expert and gallery setup for data collection. We thank Andreas Bye, of IFE, for his effort in making this study possible.

7 REFERENCES

1. K. R. Murphy and C. O. Davidshofer, *Psychological testing : principles and applications*, 4th ed. Upper Saddle River, NJ, USA: Prentice Hall, (1998).
2. N. Lau, "Situation Awareness in Process Control and the Process Overview Measure," Unpublished doctoral dissertation, Dept. of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada, (2012).
3. G. Skraaning Jr., N. Lau, R. Welch, C. Nihlwing, G. Andresen, L. H. Brevig, Ø. Veland, G. A. Jamieson, C. M. Burns, and J. Kwok, "The ecological interface design experiment (2005)," OECD Halden Reactor Project, Halden, Norway HWR-833, (2007).

4. 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).
5. G. Skraaning Jr, M. H. R. Eitrheim, N. Lau, C. Nihlwing, L. Hurlen, and T. Karlsson, "Coping with automation in future plants: Results from the 2009 HAMMLAB experiment," OECD Halden Reactor Project, Halden, Norway HWR-937, (2010).
6. T. Karlsson, H. Jokstad, B. D. Meyer, C. Nilhlwing, S. Norrman, E. K. Puska, P. Raussi, and O. Tiihonen, "OECD Halden Reactor Project: The HAMBO BWR simulator of HAMMLAB," Institutt for Energiteknikk, Halden, Norway Tech. Rep. HWR-663, Feb (2001).
7. M. H. R. Eitrheim, G. Skraaning jr., N. Lau, T. Karlsson, C. Nihlwing, M. Hoffmann, and J. E. Farbrot, "Staffing strategies in highly automated future plants: Results from the 2009 HAMMLAB experiment," OECD Halden Reactor Project, Halden, Norway HWR-938, (2010).
8. G. Skraaning Jr, M. H. R. Eitrheim, and N. Lau, "Coping with Automation in Future Plants," presented at the The 7th American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies (NPIC&HMIT), Las Vegas, NV, 2010.
9. P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bulletin,* **vol. 86**, pp. 420-428, (1979).