TEXT and rectangles in blue will NOT show on printed copy

Type final title of thesis or dissertation (M.S. and Ph.D.) below . If your title has changed since your submitted an Application for Graduate Degree, notify Graduate Office.

The Specificity of Metacognition in Nuclear Process Control: Results from a Full-Scope Nuclear Power Plant Simulator Experiment

---

A Thesis

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Master of Science

by

Name

Matthew Demas

Month degree is awarded August

Year 2015

APPROVAL SHEET

The thesis

is submitted in partial fulfillment of the requirements

for the degree of

Master of Science

Matthew Demas

AUTHOR *signature*

The thesis has been read and approved by the examining committee:

Please insert committee member names below:

Laura Barnes, PhD

Advisor

Nathan Lau, PhD (advisor)

Greg Gerling, PhD

Stephanie Guerlain, PhD

Accepted for the School of Engineering and Applied Science:

Craig H. Benson, Dean, School of Engineering and Applied Science

Month degree is awarded August

Year 2015

Print Form

Abstract

Nuclear power has been identified as an important energy source to meet carbon emissions goals set by the U.S. EPA. To continue power generation, the vast majority of nuclear plants are currently in the process of license renewal to extend operations. Consequently, modernization projects upgrading analog to digital system components are common and integrated system validation (ISV) of the control room demonstrating continued safe plant operation is necessary. Human performance represents an essential component in ensuring the safety of a nuclear power plant, and, as a result its measurement is mandated in ISV evaluations. The literature focuses significantly on individual human performance measures, but few representative empirical studies examine the psychometric properties of multiple measures in an integrated fashion. Furthermore, the literature lacks studies in industrial settings evaluating the specificity of operator metacognition. This thesis presents an empirical experiment employing a full-scope nuclear power plant simulator and recently retired operators to advance human performance measurements. The experiment evaluated the impact of scenario difficulty on workload (the Halden Task Complexity scale), expert-rated task performance (OPAS), self-rated task performance, and situation awareness (the Process Overview Measure). Further, relationships between externally verifiable measures and self-assessments provided the basis for evaluating the specificity of operator metacognition. Based on their correlations with scenario difficulty, Halden Task Complexity, OPAS, and self-rated task performance measures demonstrated basic sensitivity and validity. However, the Process Overview Measure did not correlate with scenario difficulty or other performance measures. Additionally, the results on the specificity of metacognition indicate that operators are capable of distinguishing between different aspects of their performance—overall task performance and situation awareness. The experimental method and results contribute to methodological practice and provide empirical evidence on human performance assessment for the nuclear domain.

ACKNOWLEDGEMENTS

I would like to thank the following individuals for making this thesis possible.

To Dr. Nathan Lau, thank you for taking me on as your graduate student. I enjoyed our car rides to CAER, stopping to get coffee at Trager Brothers, and our discussions about the finer things in life. The experiences I gained while working with you on this experiment and during our trips to both North Anna and Idaho National Lab have instilled a passion within me for human factors that will not easily fade.

To Dr. Carl Elks, thank you for being an advocate of mine throughout my tenure as a graduate student at UVa and for helping to expand my understanding of the relationship and interactions between humans and complex systems.

To Dr. Laura Barnes, thank you for your willingness to help me understand and interpret statistical models both as an instructor and as an advisor.

To Dr. Stephanie Guerlain and Dr. Greg Gerling, thank you for your guidance in both the classroom and as committee members on this research project.

To Bob Bailey, thank you for patience and for making this work possible.

To Mike Crist, Samuel Hughes, Mac McDade, and Henry Moore, thank you for your expertise, vision, and professionalism.

To Dr. Pamela Norris, thank you for providing funding that allowed my continual work on this project.

To Dr. Ron Boring, thank you for your discussions on control room validation, for sharing your experience with simulator experiments, and for inviting me to Idaho to observe and participate in your evaluations of digital control systems.

To Brittany Holt, thank you for all of your work shipping, printing, and keeping general order.

To Nick Soukhanov, thank you for all your work in preparing the control room for this experiment.

To Chris Bailey, thank you for helping to modify and set up the control room.

Finally, to my wife Mary, thank you for your support and for being there for me. You kept me sane and well-fed throughout my years in academia as I did my best to avoid “the real world”.

CONTENTS

1 INTRODUCTION

Increasingly strict mandates to reduce carbon emissions have necessitated a shift toward developing energy alternatives to fossil fuels (Plumer, 2014; World-Nuclear-Association, 2015). While wind and solar power are becoming competitive energy sources (World-Nuclear-Association, 2015), nuclear power is still viewed as an efficient method to achieve the 2030 reduced carbon emissions target set by the US federal government (Conca, 2014; Magill, 2015).

To address this demand, utilities have begun to request extensions to the operating life of nuclear power plants (NPPs) beyond the original 40-year licensing period. The NRC has approved over 70 license renewals in the past two decades, with 12 more currently under review and 17 more expected to apply for renewal in the upcoming years. However, plant safety is of paramount concern during the renewal process (Nuclear-Energy-Institute, 2015a).

New plant construction projects also play a role in increasing the nation's clean-energy capacity. Five new plants currently under construction are expected to become operational within the next five years, with nine others expected to begin construction sometime after 2020 (Nuclear-Energy-Institute, 2015b). Utilities must demonstrate that advanced safety and control systems are capable of safe operation prior to becoming operational.

1.1 Role of Humans in Nuclear Safety

Great emphasis must be placed on the plant safety to address the public perception that nuclear power is unsafe. This perception is largely due to rare, but high-impact accidents such as Fukushima, Japan in 2011, Chernobyl, formerly USSR in 1986, and Three

Mile Island (TMI), USA in 1979. However, nuclear power actually possesses a tremendous safety record. The TMI incident resulted in no casualties and negligible release of radiation (Conca, 2014). Kato (2014) accused the media of exaggerating the Fukushima site cleanup efforts.

One common thread connecting these incidents is the role of human workers (Meshkati, 1991; O'Hara, Higgins, Fleger, & Pieringer, 2012). For instance, cultural differences between commercial and navy NPP operations and poor control room design have been identified as primary culprits in the TMI incident; whereas, a risky safety culture encouraging a series of highly inappropriate control actions has been noted as the cause of Chernobyl (Barriere et al., 2000; Goldstein, 1986; World-Nuclear-Association, 2015).

In the shadow of TMI, the United States nuclear industry has sought to re-establish the relationship of trust with the American public. Advancements in federal regulations and the formation of self-policing bodies, such as the Institute of Nuclear Power Operations (INPO; INPO, 2014), ensure adequate human performance and ultimately public safety (Walker, 2000). A "culture of safety" now permeates the nuclear industry with a greater focus on training, procedural guidance in normal and abnormal operating states, and stringent licensing requirements (Macfarlane, 2014; Roth et al., 2010). These efforts have lead to a decrease in the number of nuclear incidents since the mid-1980's (Fader, 2009; World-Nuclear-Association, 2015).

1.2 Importance of Regulations in Demonstrating Safety to the Public and Regulating Human Performance

To ensure nuclear safety and address public concerns, the United States Nuclear Regulatory Commission (USNRC) regulates and prescribes guidance on human factors engineering (HFE) to ensure adequate human performance at all NPPs (Swaton, Neboyan, & Lederman, 1987). NUREG-0711 (O'Hara et al., 2012) describes the human factors review model outlining the regulatory evaluation criteria for the entire lifespan of their power plants. NUREG-0711 specifies four main phases of HFE over the lifespan of a nuclear power plant for licensing review: (i) Planning and Analysis, (ii) Design, (iii) Verification and Validation (V&V), and (iv) Implementation and Operation (O'Hara et al., 2012). A plant that is satisfactorily evaluated according to the process outlined in the four phases should provide "reasonable assurance of adequate protection of the public's health and safety" by "supporting plant safety and providing defense in depth" (O'Hara et al., 2012).

1.3 Verification and Validation

According to the NUREG-0711 review model, the V&V of the control room must demonstrate that the combined system including personnel, hardware and software will enable the safe operation of the plant. This phase must produce analytical and empirical evidence for licensing review that determines whether the HFE effort will provide the necessary assurance of public safety. Though costly, V&V is essential given the potential safety consequences of any nuclear event (Hamblin et al., 2013).

The NUREG-0711 review model further divides V&V into four activities: (i) "Sampling of the Operating Conditions, (ii) Design Verification, (iii) Integrated System Validation, and (iv) Human Engineering Discrepancy (HED) Resolution" (p. 73).

The first activity of V&V, Sampling of the Operating Conditions, ensures that licensees identify the environment and potential situations that may arise during the actual operation of the plant. The sampling should cover both routine personnel tasks and adverse events applicable to the verification phase of V&V. Successful completion of the sampling phase benefits the other phases by providing a comprehensive basis for further testing.

The purpose of the second activity, Verification, is to ensure that all equipment and tasks are accurately defined as specified in the Sampling of Operating Conditions phase. The identification of mis-specifications results in earlier identification and resolution of HEDs before the more costly validation stage.

The purpose of the third activity, Integrated Systems Validation (ISV), is to demonstrate that the design allows safe plant operation through performance based testing. ISV represents the first time that the engineered control room technologies will interface with the intended users. Successful ISV tests provide empirical evidence of safe plant operation and thus provide public assurance of safety.

The purpose of the fourth activity, Human Engineering Discrepancy (HED) identification and resolution, is to address design inadequacies/discrepancies uncovered during all other V&V activities. Addressing HEDs iteratively maximizes the safety of the final system.

1.4 ISV Activities

ISV is a complex process that can have significant societal implications and should deserve further research attention. Utility renewals of plant licenses necessary to balance emissions goals with energy demands will require component/system upgrades to continue operating safely. These upgraded systems must be validated. Thus, ISV is important for plant modernization. Additionally, ISV represents a complex environment where multiple parties in a modernization project convene including engineers, designers, HFE scientists as well as the end user (i.e. the operators). The challenge to demonstrate safe operation in such a complex environment with many different stakeholders makes ISV an interesting area where research can have major real-world implications.

ISV requires collecting multiple dimensions of human performance measurements in nuclear power plant (NPP) control rooms (O'Hara et al., 2012). These measures should include, but are not limited to, plant performance, task performance, situation awareness (SA), and cognitive workload. Though specific guidance on how to implement these measures is not provided, regulations require justification for the adopted measures to ensure the quality of the ISV. That is, scientifically sound measures of human performance are an important consideration in ISV testing.

ISV assessment of operator performance is a form of psychological testing. A psychological test is defined as a "measurement instrument that consists of a sample behavior…evaluated using established scoring rules" (Murphy & Davidshofer, 2001). The field of psychometrics handles theoretical considerations of psychological tests. Test

psychometric properties including validity and reliability have implications for the decisions made based on a test's results.

The property of validity describes the degree to which a measurement measures what it intends to measure. For example, a test to screen for successful job applicants should indicate how a candidate will actually perform in the workplace after being hired. A test not indicative of job performance would lack validity.

Another property is reliability. The reliability of a test describes the test's ability to achieve the same result with repeated measurement. For example, a reliable test of job candidate selection would result in the same score regardless of whether the applicant took the test at different points in time[1].

Psychological tests are used to make decisions about people's lives and careers. In NPP ISV activities, psychological tests of operator performance are used to assess control room technologies and have implications for the safe operation of a plant. In NPP evaluations, psychometrically sound tests provide important assurance of public safety by ensuring that performance evaluations actually (i.e., validity) and consistently (i.e., reliability) reflect plant safety during operation.

1.5 Objectives

To meet impending emissions goals, nuclear power is expected to provide a source of clean energy through both modernization and new construction projects. Human

[1]A detailed discussion of measurement properties is presented in Section 3.2.

performance evaluation during ISV testing of NPP control rooms helps to guarantee the continued safe operation of NPPs. In order to obtain the highest assurance of safety, it is important to utilize psychometrically sound measures of human performance in evaluations of the complex control room environments.

To address the needs present in impending control room evaluations, this thesis provides empirical evidence on the psychometric properties of multiple human performance measures collected during a full-scope NPP control room simulator experiment. In addition, this thesis supplements research on currently employed measures by examining a dimension of human performance—metacognition—that is not currently required in regulations, and thereby provides a more complete picture of operator performance in control room evaluations.

1.6 Organization of this Thesis

This thesis is organized as follows. Chapter 2 reviews the relevant literature on human performance evaluation in NPP control rooms for ISV. The goal of Chapter 2 is two-fold. The first goal is to identify a need for psychometric evaluation of multiple human performance indicators relevant for ISV testing. The second goal is to demonstrate that measures of the metacognition construct deserves further research attention and should be used in ISV evaluation. Chapter 3 discusses the research approach that will address the two research gaps identified in the literature review. Chapter 4 presents the experimental method for data collection and the hypotheses. Chapter 5 presents the data analysis and results of the multi-measure psychometric evaluation. After demonstrating the psychometric properties of multiple measures, Chapter 5 continues with statistical models

comparing metacognitive self-assessments at different levels of specificity by utilizing four psychometrically evaluated measures. Chapter 6 discusses the results in terms of their implication for ISV researchers, regulators and practitioners.

2 LITERATURE REVIEW

The purpose of this chapter is to present the state of the scientific literature on human performance measurement in NPP control rooms. This chapter outlines human performance measurements relevant in NPP control room ISV required by NUREG-0711. Further, research gaps in control room evaluations are discussed. Metacognition is identified as an important construct in nuclear power and a review of the current literature on the subject in the human factors domain is presented.

2.1 ISV Requirements

NUREG-0711 emphasizes four critical features of a quality ISV to ensure adequate HFE and ultimately public safety—(i) environment, (ii) participants, (iii) scenarios, and (iv) measurements. First, representativeness of the performance-based testing environment is important. The guidance suggests using a training simulator to ensure that the characteristics of the validation environment match the actual environment as closely as possible. Within the validation environment, everything from the interface, to the procedures, to data presentation and dynamics should accurately represent the actual operating environment. Such efforts are needed to help ensure that the experimental results maintain external and ecological validity (Skjerve & Bye, 2011), and to minimize the number of Human Engineering Discrepancies (HEDs) to ensure that the new technology can promote safe plant operation when deployed for actual use.

Second, the participants of the performance testing must also faithfully represent the user population. For instance, should NPP operators be the primary user group,

people from other positions, such as engineers should not be chosen (O'Hara et al., 2012). Otherwise, the HEDs may not represent conflicts that the actual users would encounter.

Third, the test scenarios or cases must cover a diverse range of plant operations and adverse events to provide a comprehensive evaluation of HFE and operator training (O'Hara et al., 2012).

Finally, human performance testing in the ISV of a control room design should be multi-dimensional to reflect the complex behaviors of the operators (O'Hara et al., 2012). NUREG-0711 specifies the following measures: plant performance, task performance, situation awareness, and cognitive workload. The regulations do not specify exactly how to measure these aspects of human performance, but they do require descriptions of the measures including the data collection method, frequency, methodological characteristics and performance criteria. Furthermore, NUREG-0711 states that the ISV measures should represent multiple aspects of operator performance to provide the greatest coverage and diagnosis of HEDs.

Among the aspects of sound ISV, human performance measurement is arguably the most challenging to accomplish and relevant for research. The aspects of the test environment, participants, and scenarios critical to ISV testing can usually be well defined (e.g., user group). The collected human performance measurements are the basis for the final decisions on control room adequacy, but the performance results are often difficult to pre-specify due to complex system and human interactions. Thus, human performance measurements for ISV deserve research attention.

2.1.1 Current Human Performance Measures Required in ISV

NUREG-0711 provides a detailed description of ISV methodology and discusses the four critical measures necessary for ISV studies: (i) plant performance, (ii) task performance, (iii) situation awareness, and (iv) workload. The selection of human performance measurement tools can be difficult due to the complexity of the testing environment and scenarios (e.g., Burns et al. 2008). Even after these measurement tools have been chosen, the regulations prescribe little guidance outlining the exact method to make an informed decision on control room technologies during ISV testing.

2.1.1.1 Plant Performance

NUREG-0711 identifies plant performance as the first human performance measure for an ISV and involves monitoring a plant's "functions, systems, or components" (O'Hara et al., 2012). This measure provides information about the performance of a crew as a whole (Ha, Seong, Lee, & Hong, 2007; Moracho, 1998; Norros & Nuutinen, 2005; Roth et al., 2010). Measures of plant performance can include monitoring parameter deviations relative to either steady state operation or to ranges specified by procedural guidance or expert opinion (Ha et al., 2007; Norros and Nuutinen, 2005; Roth et al., 2010). Plant measures provide an externally verifiable view of NPP operating states and operator process control, but fall short in informing the human performance characteristics of the system (O'Hara, Stubler, Higgins, & Brown, 1997). These measures are often insensitive to detriments in human performance caused by the failures in the design of the system that can be overcome by highly-trained operators (O'Hara et al., 1997). That is, even though the plant performance may be considerable acceptable, the mental state of the operator may be poor,

impairing his ability and capacity to handle future adverse events. Additionally, plant performance does not provide sufficient diagnostic information that is needed to improve the human performance characteristics of the system (Braarud & Skraaning Jr., 2006; Ha et al., 2007; O'Hara et al., 1997). Thus, other diagnostic human performance measures are required make informed decisions to improve human factors engineering.

2.1.1.2 Task Performance

Task performance is an important indicator in control room validation. Gawron (2008) defines task performance as "accomplishment of a task by a human operator or by a team of human operators." Task performance is highly dependent on the situation, and thus, many measures can quantify operator performance (see, e.g., O'Hara et al. 2012; Gawron 2008). Task performance measures often help identify errors of commission (executing incorrect actions) and errors of omission (failing to execute necessary actions). The USNRC encourages using task performance measures that match the complexity of the tasks in the testing scenario. A complete review of existing task measures is beyond the scope of this document; however, accuracy, efficiency, timing and completion of control actions have been used in previous NPP control room simulator tests to measure human performance (Chuang & Chou, 2008; Lin, Hsieh, Tsai, Yang, & Yenn, 2011; Strawn, 2010).

Measures of task performance typically defined by the time of an event and the actions of the operator may appear completely objective but often require some degree of expert interpretation to account for the specific operating context and steps within the relevant procedures or situation (O'Hara et al., 2012). In other words, the complexity of

representative experiments likely necessitates some level of subjectivity from process experts responsible for scoring operator actions in individual trials relative to a predefined set of optimal operator actions/criteria (Lau, Jamieson, Skraaning Jr., & Burns, 2008; Lau et al., 2010; Roth et al., 2010; Skraaning Jr., 1998). In addition, task measures sometimes fail to reflect cognitive demands on the operators in complex scenarios especially those used in validation testing (Burns et al., 2008).

In practice, nuclear power utilities employ operators/experts knowledgeable of the test scenarios to assess crew performance with respect to critical events during training and continuing licensing examinations of control room operators (Baron, 2014). In research settings, the Operator Performance Assessment System (OPAS) has been commonly adopted as a formal expert rating method to measure task performance, particularly at the OECD Halden Reactor Project in Norway (Skraaning Jr., 1998).

2.1.1.3 Situation Awareness

Situation Awareness (SA) commonly refers to the degree to which the operator "knows what is going on" (Endsley, 1995b). SA provides a valuable link between the scenario events and the operator's ongoing cognitive processes in validation of control room technologies (Flach, Mulder, & van Paassen, 2004; O'Hara et al., 1997). While the exact nature of SA is still contentious (Flach, 1995; Sarter & Woods, 1991; Stanton, Salmon, & Walker, 2015; van Winsen & Dekker, 2015), the relevance of the SA notion and measurements is both widely accepted (Endsley, 2015; Wickens, 2015) and required by regulation in ISV testing (O'Hara et al., 2012).

Four general methods commonly measure SA: (i) performance-based, (ii) (subjective) questionnaire-based, (iii) query/probe-based, and (iv) physiological measures.

Performance-based measures of SA involve utilizing a secondary task within a participant's primary task. Whether an operator notices some stimuli or the amount of time it takes an operator to complete an action are examples of performance-based measures of SA (Charlton & O'Brien, 2001; O'Hara et al., 1997). One possible caveat of performance-based measures is the potential confounding from the operator's experience, skill, or training (Charlton & O'Brien, 2001). Additionally, the technique may distract operators from the primary task (O'Hara et al., 1997).

Subjective measures of SA utilize questionnaires to enable operators to rate their own level of SA. Examples include SART (Taylor & Selcon, 1991), CC SART (Taylor, 1995), SA SWORD (Vidulich & Hughes, 1991), and SARS (Waag & Houck, 1994). These measures are low cost and fairly non-intrusive (Jeannot, 2000). However, these measures require that an operator be conscious of their decision making process. Further, subjective questionnaires may suffer from response biases (Jeannot, 2000; O'Hara et al., 1997).

Query/probe-based measures of operator SA involve questioning the operator directly on specific aspects of the current situation either with or without a pause during the simulation scenario trial. The most commonly used query measure is SAGAT (Endsley, 1995a). A variant of SAGAT known as SACRI (Hogg, Folleso, Strandvolden, & Torralba, 1995) was developed for use in nuclear process control. The Process Overview Measure (Lau et al., 2010) was developed to advance SAGAT and SACRI, and has been used in recent NPP simulator studies (e.g. see Burns et al. 2008). Memory effects can confound

these measures, and certain questions may cue operators to either areas relevant to upcoming events or areas unimportant to the current situation (O'Hara et al., 1997).

Physiological measures of SA collect biological data from participants during scenario trials (Charlton & O'Brien, 2001). Measures such as blink rate, heart rate and EEG have been shown to correlate with different components of SA (Vidulich, Stratton, Crabtree, & Wilson, 1994; Wilson, 2000). Additionally, eye-scanning patterns can help to identify areas where an operator looks to determine what information is being considered (Charlton & O'Brien, 2001; Droeivoldsmo et al., 1998). Physiological measurements appear objective, but require substantial interpretation for scoring SA. There are also potential confounding effects with other psychological constructs such as workload and stress. Physiological measures should be interpreted with caution and in context with other measures to ensure that they provide valid conclusions.

2.1.1.4 Workload

Measurements of operator cognitive workload form another requirement for control room design validation (O'Hara et al., 2012). Cognitive workload is defined as the level of consumption of mental resources (Charlton & O'Brien, 2001) and often correlates with operator errors, particularly in extreme situations (Wickens, Hollands, Banbury, & Parasuraman, 2013). Furthermore, even if high workload does not degrade task performance in a simulated environment/testing (as can be the case with experienced or highly trained participants), high levels of workload for long periods of time do not promote safe power plant operation (O'Hara et al., 1997).

Three methods commonly measure workload: (i) spare-capacity measures, (ii) subjective measures and (iii) physiological measures (Charlton & O'Brien, 2001; O'Hara et al., 1997).

Spare-capacity workload measures employ an artificial secondary task to evaluate the remaining mental resources not consumed by a participant's primary task (Charlton & O'Brien, 2001; Gawron, 2008; O'Hara et al., 1997). While a number of different spare-capacity measures are available (see Gawron, 2008), caution is required in a representative setting, in which the participant may resist performing artificial tasks (e.g., simple addition) due to their perceived triviality and thus hinder face validity. Additionally, the spare-capacity measure should target the same type of mental resource as the primary task according to the multiple resource theory (Tsang & Vidulich, 1989; Wickens, 2008; Wierwille & Eggemeier, 1993). Furthermore, the experimenter needs to emphasize the secondary nature of the artificial task carefully in order to avoid altering the primary task or ignoring the secondary task entirely (Lysaght, Hill, Dick, Plamondon, & Linton, 1989).

Subjective measures are often considered the assessment most faithful to the cognitive workload construct or experience (Charlton & O'Brien, 2001). A large number of subjective measures are available and applicable in a variety of domains. Examples include the Cooper-Harper Scale (Cooper & Harper Jr, 1969), the Modified Cooper-Harper Scale (Wierwille & Casali, 1983), SWAT (Reid & Nygren, 1988), and NASA TLX (Hart & Staveland, 1988). Additionally, the Halden Task Complexity Scale (Braarud, 2000) was developed specifically for applications in nuclear process control. Subjective workload measures have a number of benefits for human factors experiments. Subjective workload measures tend to

have high degrees of operator acceptance, are sensitive to changes in workload, and are diagnostic in determining the locus of the operator's workload (Hill et al., 1992; Wierwille & Eggemeier, 1993). However, response biases from participants may influence the results. At the same time, workload may not linearly relate to task performance, and there is little consensus concerning "how much workload is too much workload" in the literature (Reid & Nygren, 1988; Rueb, Vidulich, & Hassoun, 1992; Wierwille & Eggemeier, 1993). Thus, careful selection of subjective measures of workload for validation studies is important.

Physiological measures record involuntary biological responses to external stimuli. Examples include heart rate, heart rate variability, respiration rate, blink rate, pupil dilation, eye fixations, galvanic skin response and electroencephalography (EEG) (Charlton & O'Brien, 2001; Strawn, 2010). While physiological measures of workload have been employed in studies outside of laboratory settings, their properties collectively and individually are still a source of research (Matthews, Reinerman-Jones, Barber, & Abich, 2015). While physiological measures can be costly to implement and may be perceived as intrusive by participants (Cain, 2007; O'Hara et al., 1997), they avoid response biases found in self-rating measures, thus providing additional indications of workload.

2.2 Integrated Human Performance Research in Nuclear Power

The section above demonstrates a sampling of the research on individual measures of human performance and provides the basis for current regulatory guidance. This type of research is important to establish behaviors of individual measures. However, it is also important to evaluate multiple measures of human performance concurrently as nuclear process control involves complex cognitive

work that needs multiple dimensions to describe adequately (Miberg Skjerve & Bye, 2011). To date, there are few full-scope simulator experiments in the open literature that have evaluated multiple human performance measures in the context of ISV. The following section reviews NPP simulator experiments that have employed multiple human performance measures.

Lang et al. (2002) evaluated the performance of five licensed crews participating in ten scenarios in a benchmark system evaluation. The goals of the study were to establish the benchmark performance of the current control room design, determine if employing multiple versions of the same measure is meaningful, and if correlations existed between the actions taken by operators and the psychological measures of performance. The team collected both "outcome" (e.g. action appropriateness, response time, action deviations) and "process" (e.g., workload (NASA TLX), SA (SACRI, self-reports, and open ended questionnaires), teamwork (self-reports and expert ratings), and expert ratings on operator detection and diagnosis) performance measures. Lang et al. (2002) established positive relationships between the externally verifiable outcome and the more subjective process performance measures, thus providing evidence for using process measures in addition to outcome measures which are traditionally the primary performance measures employed in power utility operator evaluation. Additionally, different measures of SA revealed correlations with different aspects of performance. The authors suggest further investigation into this finding, but feel confident that the use of multiple SA measurements is important in future studies.

Roth et al. (2010) described efforts to validate a new control room for an advanced nuclear power plant currently undergoing the licensing process. In their study, the

researchers employ “multiple converging measures” to identify HEDs to address iterative improvements to the control room design by using representative scenarios with eight crews of licensed NPP operators. Expert and self-ratings on crew performance, workload and SA were employed. Records of operator actions and response times, and values of plant parameters were also used as indicators of human performance. The authors claimed that employing those measures led to successful identification of HEDs for the control room design, demonstrating practical value in multidimensional measurements of human performance. However, the authors do not present exact method of data collection (e.g., questionnaires) and empirical results for the study (Mitsubishi Heavy Industries Ltd, 2008). Further, the authors relied only upon discussion to establish convergence of measurement results. Thus, this study provides limited support for other researchers and practitioners in the evaluation of multidimensional human performance assessment in ISV.

Dong et al. (2013) expressed the importance of integrating timeline data in their validation observations of a crew operating an advanced control room. They employed measures of usability (e.g. interface actions, mouse clicks, etc.) along with subjective measures of SA and workload. However, they only presented the results of the usability measures and omitted the discussion of workload and SA. Their work contributes to the importance of uncovering HEDs during the stages of V&V and demonstrates a method in which this can be accomplished. However, the limited scope of the analysis in their article offers limited supported for other practitioners conducting ISV.

The Human System Simulator Laboratory (HSSL) at Idaho National Laboratory (INL) focused on the usability aspects of evaluation as the initial step towards aiding utilities in selecting appropriate measures for ISV (Boring, Lew, Ulrich, & Joe, 2014; Le Blanc, Boring,

& Gertman, 2001; Ulrich et al., 2012). A single crew of operators participated in walk-throughs of simulated malfunction scenarios using the current and proposed implementations of a digital turbine control system (TCS). During the scenarios, subjective feedback, expert review, simulation and behavioral logs were collected in order to inform design changes for future versions. The subjective feedback and expert review were stated to provide immediate benefit, but the simulation and behavioral logs were not analyzed and remained untapped potential. The primary contribution of this work is the demonstration of an evaluation method for early stage developments of control room technologies. For the same reason, the formality in data collection did not (and need not) meet the standard of final design validation. Further, the authors recognize that the walk-through nature of the scenarios used with the upgraded TCS hinders the analytical results of the work, thus the conclusions drawn are limited with regards to meeting full-scale validation criteria outlined in NUREG-0711.

2.2.1 Summary: Current Research Needs with Existing Measures

Significant research has been directed at advancing human performance measurements in the nuclear domain but the literature contains virtually no detailed empirical investigations on the measurement properties of multiple human performance measures in representative nuclear process control settings (c.f., Lang, et al., 2002). Given that the outcome of ISV depends significantly on collecting multidimensional human performance data in simulator studies, psychometric investigation of commonly employed and novel measurement instruments deserves continual research attention.

The quality of ISV depends in part on using psychometrically sound measures. Despite this importance, few studies published have utilized multiple human performance measures and presented their psychometric properties in the literature. This thesis is of great practical importance for the nuclear community including regulators and utilities, because empirical evidence on the interactions and relationships between human performance measures help draw conclusions in ISV studies based on their measurement properties.

Given the above attempts at validation of new technologies in the NPP control room environment, there is a need for the demonstration of a variety of novel human performance measures in a high fidelity control room environment that improves and expands the knowledge of NPP operator performance. However, there have been relatively few studies that have focused on implementing additional human performance measures novel to the nuclear process control.

2.3 Advancing and Evaluating New Human Performance Measurement

The research community is continually advancing and evaluating human performance measurements to support ISV of NPP control rooms. The OECD Halden Reactor Project (HRP) has been most active in NPP control room research (Miberg Skjerve & Bye, 2011). They have advanced a wide range of research topics including teamwork, human-automation interaction, adverse scenarios, and human reliability analysis as well as human performance (Braarud & Svengren, 2006; Broberg, Hildebrandt, Massaiu, & Braarud, 2008; Bye et al., 2011; Hallbert, 1997; Lau, Skraaning, Jamieson, & Burns, 2008; Skjerve, Nihlwing, Nystad, & Strand, 2009; Skjerve & Strand, 2007; Skraaning Jr. & Miberg

Skjerve, 2006; Skraaning Jr., Eitrheim, & Lau, 2010). However, HRP research is mainly documented in internal reports rather than open literature.

The open literature contains a series of publications on simulator studies using student participants to advance human performance measurements in the context of nuclear process control. The University of Central Florida used ECG and EEG to measure the workload of university students operating a full-scope NPP simulator (Guznov, Reinerman-Jones, & Marble, 2012; Leis, Reinerman-Jones, Mercado, Barber, & Sollins, 2014; Mercado, Reinerman-Jones, Barber, & Leis, 2014; Reinerman-Jones, Guznov, Mercado, & D'Agostino, 2013). Researchers working with the Institute of Nuclear Energy Research in Taiwan have also published a series of low-fidelity simulator experiments employing university students assessed by traditional human performance measures and physiological instruments. The research results contribute to team workload assessment techniques and models of secondary task performance using multiple physiological indicators (Hwang et al., 2008; Lin et al., 2011). Measures of individual human performance dimensions are a continual focus in non-representative/low-realism research for the nuclear domain.

The Korea Advanced Institute of Science and Technology (KAIST) has conducted experiments with university students and professional operators to advance their comprehensive measurement system known as the Human Performance Evaluation Support System (HUPESS; Ha & Seong, 2009). Their experiments with student participants led to a method for measuring operator information processing capacities using eye-tracking (e.g., J. T. Kim, Shin, Kim, & Seong, 2013) while their studies with expert operator participants advanced measures of task performance using plant performance (Jang, Park,

& Seong, 2012) and communication classification methods (A. R. Kim et al., 2012; A. R. Kim, Lee, Park, Kang, & Seong, 2013). KAIST continues to extend human performance evaluation methods.

Researchers continue to focus on advancing and developing new measures of human performance in NPP control rooms to provide the most complete assessments and promote the greatest level of safety. This thesis seeks to continue the goal of measure advancement through the conducting of a full-scope simulator experiment (Demas, Lau, & Elks, 2015).

2.4 Metacognition: A New Measurement Construct for Nuclear Process Control

Current regulatory guidance characterizes operator performance through the highly established measures of plant performance, task performance, situation awareness, and workload. The studies discussed above demonstrate that advancements in measurements can further elucidate understanding of human performance in control room evaluations. Metacognition is one construct currently in development that has attracted attention in a variety of domains including aviation, off-shore drilling, driving and nuclear power (Lau, Skraaning Jr, & Jamieson, 2009; J. D. Lee, 1999; Roberts, Flin, & Cleland, 2014; Sulistyawati, Wickens, & Chui, 2011). While not currently required by regulation, metacognition has the potential to improve ISV. Thus, metacognition research could prove useful in diagnosing and evaluating future NPP control room technologies. The following section presents the relevant aspects of metacognition in the domain of human factors for nuclear process control.

2.4.1 General Characterization

Metacognition describes "the experiences and knowledge (people) have about (their) own cognitive processes" (Schwartz & Perfect, 2002). Existing models of metacognition distinguish between knowledge of and regulation of cognition (Schraw, 1998; Schraw, 2009).

Schraw (1998) distinguished between the subtypes of the regulation of cognition into planning, monitoring and evaluation. *Planning* refers to the appraising of strategies and demands prior to attempting a task. *Monitoring* is defined as "one's on-line awareness of comprehension and task performance" (ibid, p. 115). Finally, *evaluation* occurs after a task and addresses "the products and efficiency of one's" performance (ibid, p. 115).

Logically separate from metacognitive regulation is metacognitive knowledge, which describes "what individuals know about their own cognition or about cognition in general" (ibid, p. 114). Schraw further classified knowledge of cognition into three types: declarative, procedural and conditional. Metacognitive *declarative* knowledge describes understanding of one's own cognitive processes and "what factors influence one's performance" (ibid, p. 114). *Procedural* knowledge, on the other hand, describes one's ability to know "'how' to do things" (ibid, p. 114), while *conditional* knowledge describes "when and why to use declarative and procedural knowledge" (ibid, p. 114).

One method to assess metacognition is through metacognitive judgments that require a subject to provide an externally verifiable self-assessment. These evaluations can examine the content and regulatory aspects of operator metacognition (Schraw, 2009).

Metacognitive judgments can be of either a global (referencing a general aspect) or item-specific nature. Furthermore, item-specific metacognitive judgments possess a number of characteristics that categorize the various aspects of a participant's response. These characteristics include absolute accuracy (calibration) and relative accuracy, bias, and discrimination (resolution; Schraw, 2009).

Absolute accuracy (or calibration) refers to the "discrepancy between a confidence judgment and performance" (Schraw, 2009). Calibration is most often assessed using the absolute accuracy index, which is a term from a decomposition of the Brier score (Baranski & Petrusic, 1994; Schraw, 2009). Relative accuracy, on the other hand, refers to the "relationship between a set of confidence judgments and performance scores" (Schraw, 2009). Relative accuracy is typically assessed using correlations between actual and perceived performance. Metacognitive bias represents over- or under-confidence in perceived judgment accuracy. Bias is typically assessed through the bias index (Lichacz, 2008; Schoenherr, Leth-Steensen, & Petrusic, 2010; Schraw, 2009). Finally, discrimination (or resolution) refers to "the degree to which an individual distinguishes between confidence judgments for correct versus incorrect [assessments]" (Schraw, 2009) and is typically measured by the discrimination index that, like calibration, is also a term in a decomposition of the Brier score (Baranski & Petrusic, 1994).

Another characteristic of human judgments, specificity, has been addressed by researchers studying trust in automation (Lee & See, 2004). Specificity is defined as "the degree of to which [a characteristic] is associated with a particular component of the [characteristic holder]" (Lee & See, 2004). Specificity has received some attention in the context of metacognition in the fields of education and criminal justice through the study of

"grain-size" in certain recall situations (Dunlosky, Rawson, & Middleton, 2005; Krug, 2007; Schraw, 2009). However, no studies have addressed grain-size from a human performance perspective in industrial work settings.

The characteristics of calibration, resolution, and specificity have important implications especially in safety critical domains such as air traffic control, command and control, and process control. In these domains, deficiencies in any one of these characteristics could result in unsafe actions, reduced team effectiveness, or communication failures that ultimately may endanger the safety of the public or result in the loss of human life. For instance, poor calibration is associated with systematic deviations of confidence-accuracy assessments from ideal and can result from either over-confidence or under-confidence. Under-confident operators that are may be unwilling to seek action when action is needed, while over-confident operators may act inappropriately when no action or another action would be preferable. Poor resolution on the other hand may manifest in the inability to determine the exact level of one's confidence. Operators with poor resolution may reduce the efficiency of teams in cooperative settings in cases where they appear confident, but actually do not have correct assumptions about a situation. Thus, an operator's inability to resolve his level of confidence may reduce the amount of trust that other team members have in that operator. Finally, poor specificity may result in the inability to determine different levels of one's performance from one another or to differentiate between different aspects of a situation. Poor specificity might result in over generalizing aspects of a situation that could lead to failures in communicating the necessary information to other team members. Such breakdowns in communications and decision-making resulting from poor calibration, resolution and

specificity have the potential to cause catastrophic failures in safety critical domains. Thus, the study of these characteristics deserves significant research attention.

2.4.2 Metacognitive Judgments in Human Factors

Self-assessment relating to one's own performance (Lichacz, 2008; Schoenherr et al., 2010; Schraw, 2009) has been studied in myriad domains including driving, command and control (C2), aviation, air traffic control, and nuclear power (Lau et al., 2009; Lee, 1999; Roberts et al., 2014; Sulistyawati et al., 2011). At the core of these studies is the connection between a self-assessments and externally verifiable assessments. While the vast majority of the literature has devoted attention to the characteristics such as calibration and resolution, virtually no studies have addressed the specificity of metacognitive judgments. Evaluations of this type could be useful in NPP evaluations such as ISV and to the human factors community in general.

Lee (1999) investigated the effects of different information presentations on objective SA and SA confidence of participants in a simulated driving task. The relative accuracy of the participants (as assessed through correlations between SA and SA confidence) was found to be low ($r(16^*)$=0.11 and $r(16^*)$=0.21 for older and younger drivers, respectively). The authors graphically assessed metacognitive bias by plotting averaged correctness versus averaged confidence. These plots revealed that older drivers tend to be over-confident, and younger driver drivers tend to be under-confident.

Lichacz, Cain and Patel (2003) examined the effects of task demands on SA, SA confidence, and metacognitive bias in novice participants engaged in a simulated air traffic control task. Metacognitive bias was calculated as the difference between average SA

confidence and average SA correctness. The authors found an inverse relationship using repeated-measures ANOVAs between task demands and participant confidence, but no relationship on confidence bias. The authors noted that less and more complex queries resulted in under- and over-confidence, respectively.

McGuinness (2004) evaluated SA and SA confidence using QUASA of experienced participants in a simulated C2 task. SA queries were distributed at regular intervals and the data were analyzed using signal detection theory. Metacognitive bias scores were examined graphically as the difference between average confidence and average accuracy for each team in the experiment. The authors found that utilizing multiple assessments of response tendencies (both implicit and self-assessed) revealed greater diagnostic information about team differences extending beyond simple SA accuracy.

Lichacz and Farrell (2005) examined the effects of different scenarios on SA and SA confidence of experienced participants from 5 different counties in an operational net assessment experiment. Using answers from SAGAT queries and confidence ratings, the authors examined relative accuracy through correlations and resolution through an ANOVA by using the confidence level as a factor. The authors determined that confidence level was related to accuracy ($F(4,180)=40.32$, $p<0.001$). They concluded that objective SA (SAGAT proportion correct) and subjective assessments on queries (confidence ratings) agreed with findings by Lee (1999) and Licacz et al. (2003).

Lichacz (2008) evaluated SA and SA confidence of experienced military and civilian government planners and system of systems analysts participating in a simulated C2 task. Using the calibration analysis framework, the author computed calibration, resolution, and

bias scores (Baranski & Petrusic, 1994; Schoenherr et al., 2010) as well as relative metacognitive accuracy (correlations). The author found a decrease in both SA and SA confidence over probe sessions. However, the resolution of SA confidence with respect to their actual SA increased over the sessions, indicating that the participants were better at distinguishing (or more sensitive to) their confidence level with respect to their SA level. The author also concluded that the calibration analysis framework provides a useful tool to examine the relationship between SA and SA confidence while separate analyses on individual measures could obscure the results.

Lichacz (2009) evaluated SA, SA confidence and sleepiness in response to cultural differences in a multinational coalition operation experiment designed to study strategic planning processes. As in Lichacz (2008), the author used the calibration analysis framework to compute calibration, resolution, bias scores and relative metacognitive accuracy (correlations). Both different nationalities and planning strategies led to different confidence biases. SA scores were inversely related to SA confidence, with higher SA coupled with under-confidence and lower SA coupled with over-confidence. This study contained no other measures of decision-making or task performance for comparison.

Rousseau et al. (2010) evaluated correlations between a SA measure with objective and subjective components (QUASA; Edgar, Edgar & Curry, 2003) and a subjective measure of SA (the 3-D version of SART; Taylor, 1990) in a simulated C2 counter-terrorism planning task. Additionally, the authors examined the QUASA accuracy/confidence ratings in terms of relative accuracy (correlation statistics), graphic calibration, and metacognitive bias. The authors report negative correlations between the QUASA accuracy scores and subscales of SART ($r=-0.438$ for supply, $r=-0.363$ for understanding, and $r=-0.569$ for global), indicating

dissociation between the subjective and objective measures of SA. Additionally, the authors observed a positive correlation (r=0.458, p=0.04) between the QUASA confidence scores and SART Supply (S) scores.

Sulistyawati, Wickens and Chui (2011) assessed the relationship between metacognitive bias, Endsley's three levels of SA (Endsley, 1988, 1995a) and task performance of experienced pilots using multiple linear regression. The authors calculated metacognitive bias as the difference between the averaged binary confidence score and the average SA score (proportion correct responses). More difficult SA queries (associated with Levels 1 and 3 in this particular experiment) were associated with greater overconfidence, in agreement with previous studies (Lichtenstein, Fischhoff, & Phillips, 1982). The multivariate linear regression model indicated an inverse relationship between performance and confidence bias ($F(4,11) = 3.58, p < 0.05$; $Adj.R^2 = 0.41$; $\beta_{overconf} = 0.532, p = 0.039$). This relationship indicates that pilot overconfidence is associated with low task performance. However, the use of bias scores where zero represents ideal calibration is difficult to interpret (Lau, Skraaning Jr, & Jamieson, 2009). The reason for this difficulty is that both over-confidence (positive bias) and under-confidence (negative bias) represent undesirable states. However, the tendency for a data set to be largely positive or negative allows for this aspect of the interpretation to be relaxed.

2.4.3 Assessing Metacognition (in the Control Room)

Metacognition has received limited attention in the nuclear domain, with only two previous studies on measuring the construct (Lau et al., 2009; Skraaning Jr. & Miberg

Skjerve, 2006). Both studies measured metacognition based on the difference between operator performance and self-assessed performance.

Skraaning and Skjerve (2006) evaluated metacognitive accuracy and trust in automation of operators in knowledge-based (non-procedurally guided) and rule-based (procedurally guided) process malfunction scenarios in a full-scope NPP control room simulator experiment. Metacognitive accuracy is calculated as the difference between standardized plant performance (see 2.1.1.1) and standardized, averaged self-rated crew performance. Trust in automation was highly correlated with metacognitive accuracy in knowledge-based scenarios, but uncorrelated in rule-based scenarios. The authors speculated that the procedural guidance in rule-based scenarios promoted properly calibrated metacognitive accuracy by clearly delineating the actions to be performed by members of the crew and by the automation. Further, without procedural guidance in knowledge-based scenarios, the crew may have included failures of the automation in their self-ratings of performance. This study utilized a metacognitive accuracy measure based on overall performance and self-rated data were aggregated to the plant performance and crew level to achieve the comparison. Thus, this study focused on a different level of specificity from SA-based metacognitive performance measures reviewed earlier.

Lau, Jamieson and Skraaning (2009) evaluated the effects of scenario type on metacognitive bias and accuracy in a full-scope NPP control room simulator experiment. Metacognitive bias was calculated as the difference between standardized scores of self-rated and expert-rated performance. Metacognitive accuracy was calculated "by i) calculating the root-mean-square (rms) of the Metacognitive bias scores, and ii) subtracting each (rms) score from the maximum score" (Lau et al., 2009). Knowledge-based scenarios

led to significantly lower metacognitive accuracy scores than procedural-based scenarios, and metacognitive bias negatively correlated with workload. The authors discussed their findings within the context of supporting training and the designing of appropriate control room technologies. Similar to the Skranning & Skjerve (2006) study, the metacognitive bias and accuracy measures focused on the overall task performance level of specificity.

2.4.4 Summary of Metacognitive Judgments in Human Factors

The existing literature on SA metacognitive judgments has revealed several important findings. First, correlations between overall confidence and overall accuracy are generally low or non-existent. This tendency corroborates findings from other domains (Dunlosky et al., 2005), and suggests that the construct should be studied with more than mere correlations.

Second, analyzing metacognitive judgments by connecting confidence and accuracy reveals information not accessible by examining either measure alone. Most researchers have examined metacognitive bias with either correlations or graphical inspection. Additionally, some investigations into the SA accuracy-confidence relationship examine participant accuracy conditioned upon confidence level (Lichacz & Farrell, 2005; Lichacz, 2008; Lichacz, 2009; Rousseau et al., 2010). However, these studies only feature assessments at a single level of specificity.

Third, only Rousseau et al. (2010) made comparisons between higher specificity assessments (with SA accuracy and SA confidence on individual queries) to lower specificity assessments (with a self-assessment of overall SA). The results of this analysis

suggest further investigations are necessary. In brief, the literature lacks studies on the relationship between metacognitive judgments at varying levels of specificity.

While the above studies demonstrate the sensitivity of crew-level operator metacognitive accuracy to differences in scenario type, there are currently no studies which address the relationship between measures of metacognition at differing levels of specificity in a single NPP control room experiment. It is unknown if operators are sensitive to their performance at different levels of specificity. Thus, this subject deserves further research attention.

2.4.5 The Importance of Evaluating Metacognitive Specificity

Research on the specificity of self-assessments is important because of the potential applications to improving ISV evaluations of automated systems. Automation will become increasingly widespread as plants renew their licenses and must modernize aging systems with advanced control systems (Boring et al., 2014). Appropriate use of automation promotes safe plant operation (Hoff & Bashir, 2015) and must be addressed in ISV evaluations of automated systems.

Appropriate usage of automation has been linked to operator trust; factors affecting trust have received considerable attention in the literature (Hoff & Bashir, 2015; Lee & See, 2004). Studies have revealed that operator usage of automation involves a comparison between self-assessed ability and an assessment of the automation's ability (de Vries, Midden, & Bouwhuis, 2003; Lee & Moray, 1994). Operators decide to use automation when their trust in the automated system outweighs trust in their own abilities. Additionally, studies have further identified that operators can be specific in their assessments of

automation and can correctly associate failures in functionally distinct aspects of automation (Lee & Moray, 1994; Lee & Moray, 1992; Muir & Moray, 1996). However, further studies into the specificity of *operator* self-assessments are lacking.

Thus, evaluations into the specificity of operator self-assessments provide an important source of research for both the human factors community in general and those researchers studying human-automation interaction.

2.5 Summary

This chapter began with a discussion on the human performance measures currently required in ISV evaluations, and the need for ISV studies including psychometric evaluations of multiple human performance measures was demonstrated. Additionally, the motivation for including novel human performance measures was discussed, and metacognition was introduced as a construct for future NPP control room evaluations. The specificity of metacognition was identified as an important topic with practical connections to the evaluation of automation. The next chapter outlines the research approach to be taken to address to address the dearth of literature concerning psychometric evaluations of multiple human performance measures and the advancement of human performance assessment through the evaluation of the specificity of metacognition.

3 RESEARCH APPROACH

The current literature indicates two significant research gaps for ISV activities. First, relatively few representative full-scope NPP control room experiments with expert participants have been conducted in recent years with results published in the open literature. Within that set of publications, virtually none present psychometric evaluations of multiple human performance measures in the context of NPP control room ISV. Without knowing the psychometric properties, conclusions made with human performance measures are difficult to assess. Thus, psychometric properties of required human performance measures are important for ISV and promoting plant safety.

Second, the literature has demonstrated the importance of metacognition in both observational and experimental NPP control room studies (Carvalho, dos Santos, & Vidal, 2006; Lau et al., 2009; Vicente, Mumaw, & Roth, 2004). Researchers in domains such as air traffic control and command-and-control have demonstrated sophisticated analysis techniques and highlighted metacognition's role in decision-making. However, the literature has not attended to the specificity of metacognition in NPP operator self-assessments. New knowledge and measurements of metacognition have the potential to improve the assurance of safe operation. However, due to the limited number of studies in the NPP domain, the regulations do not currently mandate the construct.

3.1 Research Objectives

This thesis will address the two research gaps in the literature by accomplishing the following two objectives to promote safe plant operation in NPPs. The first objective is to collect data on multiple human performance measures and evaluate their psychometric

properties, providing ISV practitioners and regulators a reference data set to guide future ISV activities. Measures that vary as expected and possess adequate psychometric properties demonstrate strong potential for future adoption in ISV activities.

The second objective is to provide empirical evidence on the specificity of operator metacognition by collecting data to demonstrate the merit of metacognitive assessments at different levels of specificity. Metacognitive judgments made by operators at different levels of specificity should be more indicative of their performance at each level, respectively.

3.2 Full-Scope Simulator Human-in-the-Loop Experiment

To address the aforementioned objectives and provide (1) empirical evidence on the properties of multiple human performance measurements and (2) an investigation into the specificity of metacognition, a full-scope human-in-the-loop NPP control room experiment was conducted. In this experiment, expert participants were recruited to perform representative tasks in a representative environment. Such lengths were required to ensure a high level of representativeness and provide the greatest possibility of the results generalizing to actual ISV evaluations.

3.2.1 Evaluation of Measurement Properties

Human performance measurements must possess qualities of reliability and validity to ensure generalizability. Reliability ensures that repeated application of the measure yields consistent results, while validity ensures that the measure assesses the targeted construct so interpretation of the results is accurate.

This study tested reliability of measurements by assessing the internal consistency of questionnaires with Cronbach's alpha (Cronbach, 1951). Internal consistency measures describe the degree to which questionnaire item scores are similar to scores of other items on the same questionnaire. Higher internal consistency scores indicate that the construct operationalized in a questionnaire is reliably measured.

This study also tested the inter-rater reliability of scores assessed by multiple experts with intra-class correlations (ICC; Shrout & Fleiss, 1979). Inter-rater reliability statistics evaluate the degree of consistency both between and within raters. Higher inter-rater reliability statistics indicate a greater degree of rater interchangeability. Additionally, larger inter-rater reliability scores suggest that assessment by different expert raters do not drastically change the scores obtained.

This study maintained face and content validity of the experimental environment. Face validity is defined as the degree to which an experimental aspect "appears to provide a reasonable and an acceptable" representation of the actual aspect (Murphy & Davidshofer, 2001). This experiment ensured face validity by maintaining the protocols and mannerisms employed in NPP simulator training and evaluation exercises. Domain experts were recruited to act as simulator operators and provide guidance on the experimental protocols. By maintaining these protocols, the experiment achieved participant acceptance as confirmed during participant debriefing.

Content validity describes the degree to which the content of a test covers the construct of interest. For this thesis, the construct of interest is human performance in nuclear process control. In this experiment, content validity were assured by utilizing

process expert knowledge to create realistic experimental scenarios, environment, and tasks that resemble those found either in actual NPP simulator exercises or in actual operation. Content validity were also assured by utilizing measurements required by ISV testing to maintain assessments that would be found in an actual ISV test. Formal assessment of content validity was not conducted, as the requirement on domain knowledge would ultimately lead to higher reliance on expert opinions. Increased content validity ensures greater generalization of the experimental results.

The efforts to ensure face and content validity provide a preliminary basis for construct validity of human performance assessment. Construct validity can be examined through the relationships between multiple measures to evaluate whether they vary according to expectation. When multiple measures vary as expected, convergent validity is achieved. Standard cutoffs for correlation strengths should be used to assess the relationships between two measures and the constructs they represent. The assessment of construct validity comprises two aspects—expected correlations and observed correlations. When expected correlations are low or non-existent and the actual correlation is close to zero, discriminate validity is achieved. However, when a correlation is expected for convergent validity, the actual correlation should not be too large. Correlations close to ±1.0 indicate that the measures may be actually assessing the same rather than different constructs.

By addressing reliability and validity, the effectiveness of measures may be assessed and conclusions can be made from the experimental data in the light of the measurement properties.

3.2.2 Specificity of Metacognition

To evaluate the specificity of operator metacognition, comparisons were made between externally verifiable measures and self-assessments. In the experiment, two externally verifiable measures—task performance and situation awareness were collected along with corresponding self-assessments. The task performance self-assessment was of a low level of specificity, while the situation awareness self-assessment was of a higher level of specificity. By creating models with the externally verifiable statements as the response variables and the self-assessments as the predictor variables, comparisons can be made to determine whether or not more specific self-assessments merely reflect the general feeling of performance. By collecting this data within the context of a representative experiment, the greatest level of generalizability is ensured.

3.3 Research Significance and Practical Implications

This thesis contributes to the open scientific literature on conducting ISV of NPP control rooms and thereby has the potential to impact safety and productivity of nuclear power generation. The full-scope simulator experiment provides empirical data on the basic reliability and validity of human performance measures commonly used in NPP simulator studies. The results of the analysis should provide a reference data set for regulators and practitioners concerned with ISV. Thus, the results have implications for ISV activities that provide public assurance of safety and play a large role in current modernization and new construction projects. This experimental data set and psychometric results also support researchers in modeling human performance in complex environments.

The examination on metacognition provides new empirical knowledge about specificity of metacognition for the scientific community and a technical basis for measurement prescription in regulations. In particular, the experiment speaks to the value of operator self-assessment that regulators and ISV practitioners may find vital in the evaluation of new technologies in control room environments.

4 METHOD

This section describes the methodological details of the full-scope simulator experiment conducted to address the current research needs outlined in Chapters 2 and 3 (Demas et al., 2015). The section is organized as follows: experimental environment, participants, procedure and experimental design, human performance measures, and hypotheses.

4.1 Experiment Environment

This experiment utilized a full-scope, Generic Pressurized Water Reactor (GPWR) simulator[2] in the control room facility at the Center for Advanced Engineering and Research (CAER)[3], Bedford, VA (Figure 1; Demas et al., 2015). The hard-wired panel user interface of the GPWR simulator was displayed across 48 24-inch monitors. Mice and keyboards were used to control the simulated process. The control room also included a supervisor's workstation that had two touchscreen monitors and housed both paper-based and digital (PDF) procedures.

Experimenters observed operators (i.e., participants) interacting with the simulator unobtrusively in the observation gallery located at the back of the control room, enclosed by one-way mirrors (Figure 2). The observation gallery housed all data collection equipment and contained a raised platform to support observation. The Noldus Observer

---

[2] http://www.gses.com/products/gpwr-nuclear

[3] CAER is located in Forest, VA and online at http://caer.us/ .

XT[4] software suite was used to integrate multi-channel audio, multi-angle video, and annotations of operator behaviors, physiological data, and plant simulator logs.

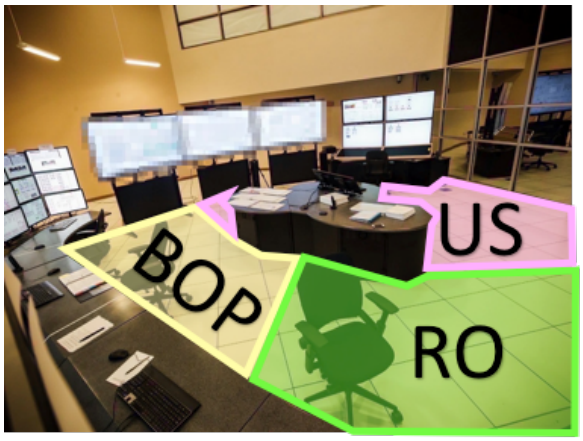

Figure 1 Overview of the CAER control room with locations of each role labeled.

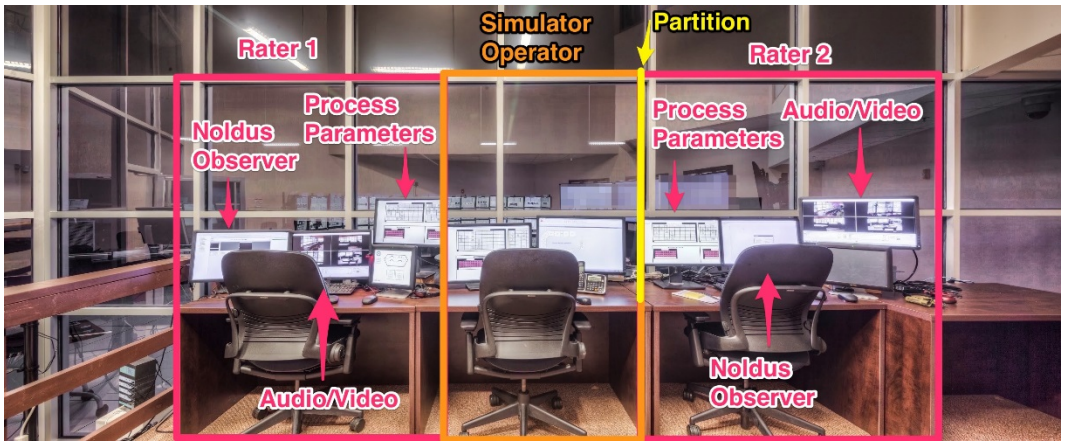

Figure 2 Figure displaying the position of experimental team members in the observation gallery.

4.2 Participants

Nine previously licensed operators (n=9) were recruited to form three crews of three members. Members of each crew were assigned the positions of unit supervisor (US),

[4] http://www.noldus.com/human-behavior-research/products/the-observer-xt

reactor-side operator (RO) and balance-of-plant-side operator (BOP). Each participant maintained the assigned position for the entire data collection period. See Figure 1 for the area of the control room generally occupied by each operator role.

4.3 Procedure

Each crew participated in the experiment for five days. A retired NPP operator trainer familiarized operators with the GPWR plant systems and conducted guided practice on operating the GPWR during the first day-and-a-half (approximately 12 hours).

After training on the morning of the second day, the operators completed a practice scenario to become accustomed to the experimental trial protocol (e.g., responding to questionnaires, wearing physiological gear, etc.). After this practice trial, the data collection trials began, continuing on the third through fifth days of the experiment.

For the data collection trials, all operators were first outfitted with data collection gear, which included wireless microphones (BOP, RO, US), Tobii™ Eyetracking Glasses (BOP, US), BioPac BN-PPGED electrodermal activity transmitters (RO, US), and BioPac BN-RESP-XDCR thoracic expansion (i.e., breathing) transducers (RO, US). Then, the operators were brought to the control room area where they received a quick briefing of the initial conditions of the scenario. The participants were asked to act as if they were on duty by resolving process disturbances to maintain plant safety and productivity for ten different scenarios.

The recently retired NPP trainer designed five scenarios with the simulated plant operating at 50% power and another five scenarios at 100% power. Each experimental scenario trial comprised two to four malfunction events and was subdivided into two

periods. The scenario designer also predefined a time for a "scenario-freeze" that signaled the end of a scenario period. During this scenario-freeze, the operators responded to human performance questionnaires at workstations away from the control room area (Figure 3).

Each scenario trial lasted approximately 1.5 hours with breaks of twenty minutes (minimum) between trials. On the fifth day, experimenters debriefed and collected feedback from the operator crew.

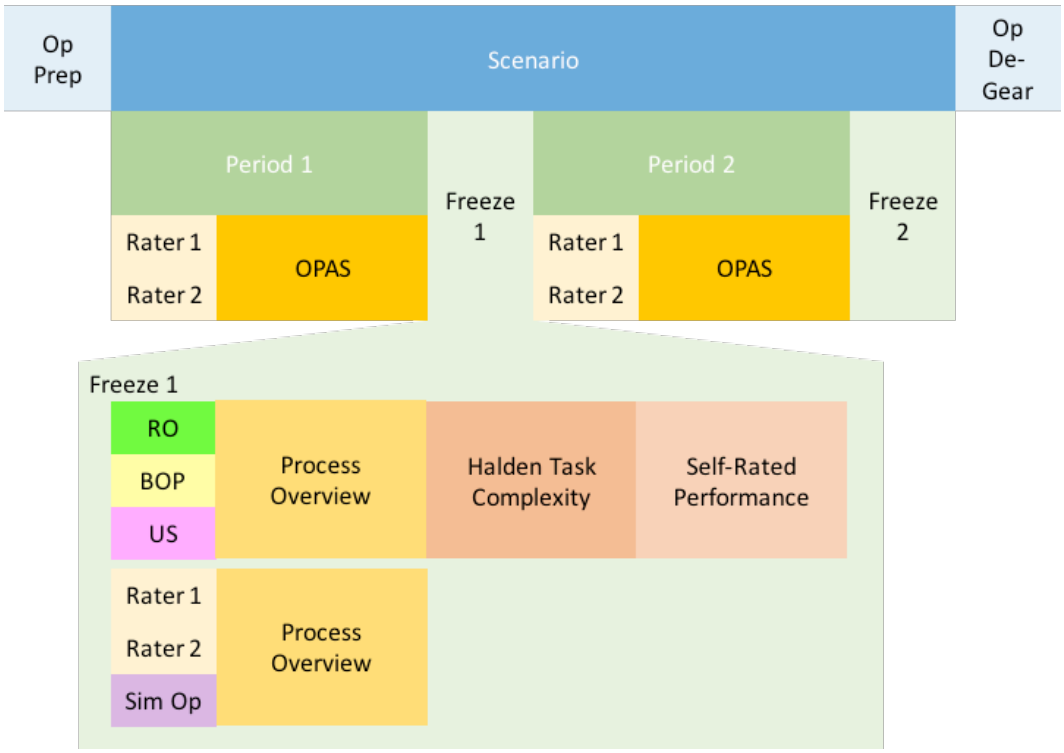

Figure 3 Depiction of scenario structure and questionnaire administration.

4.4 Experimental Design

The experiment was a within-subjects design that included a two-level factor of scenario period (first and second). Each period featured between one and three malfunctions of varying difficulties.

Two process experts independently rated every malfunction event in each scenario period on four dimensions of difficulty (detection, diagnosis, intervention, and restoration). Each dimension was rated on a five-point Likert scale. Prior to experimental data collection, the process experts pilot tested and rated the scenarios for one week. These experts were not involved in designing the scenarios. Human performance measures were examined with respect to these *a priori* event difficulty ratings to assess measurement properties.

Difficulty indices per malfunction event were averaged between the ratings of the two process experts and difficulty ratings of the events were aggregated to obtain a scenario period difficulty index per scenario period.

4.5 Human Performance Measures

This experiment employed the following measures: plant performance (based on simulator logs including alarms, operator actions and trend graphs), task performance (both expert- and self-ratings), workload, situation awareness (SA), SA confidence, and physiological measures (i.e., electrodermal activity, thoracic expansion, and eye tracking). For the purpose of examining psychometrics and metacognitive specificity, this thesis focuses on the measures of task performance, situation awareness and workload. Thus, the treatment and analysis of physiological measures is beyond the scope of this thesis.

4.5.1 Workload

The Halden Task Complexity scale (Braarud, 2000, 2001) was adapted into American NPP operator English to measure workload (see Appendix C). The Halden Task Complexity scale contained 5 items on a 7-point Likert scale and was administered during

scenario-freezes after each period. Each operator's score per scenario period was the averaged ratings of the five items. This averaged rating of the Halden Task Complexity scale corresponds to the amount of cognitive load imposed upon operators as they perform their tasks.

4.5.2 Expert-Rated Task Performance

Task performance was measured using scenario-specific rating sheets developed according to the Operator Performance Assessment System (OPAS; Skraaning Jr. 1998; see Appendix D). The scenario designer developed the OPAS rating sheets that contained performance items corresponding to individual steps necessary to resolve malfunctions. Each item included predefined performance criteria associated with a score between zero and three. Zero represented failure to complete the step and three represented an ideal response.

Two process experts (retired, formerly licensed NPP operators) rated the participants on the OPAS items from the observation gallery during scenario trials. Rater 1 was present for all three weeks of the experiment, while Rater 2 was present for weeks 1 and 3. The raters were separated by a partition in the gallery and refrained from discussion to minimize mutual rating influence. Raters also used integrated audio/video information from Noldus Observer XT and process parameter values from the simulator. Both raters received an identical audio and video feed through headphones and mirrored monitors. The data from the first and third weeks of data collection were used to generate inter-rater reliability statistics.

Experts rated approximately 10 items per scenario period. The OPAS scores were calculated for each crew per scenario period by averaging the item scores provided by all raters.

4.5.3 Self-Rated Performance

The self-rated performance questionnaire in (Skraaning Jr. et al., 2007) was adapted into US NPP operator English to measure operator self-assessed task performance (see Appendix E). The self-rated performance questionnaire contained 10 items on a 5-point Likert scale and was administered during scenario-freezes after each period. Prior to computing the scores, the ratings of the tenth item were inverted. This inversion was necessary to account for the fact that higher scores on the 10^{th} item corresponded to poorer self-rated performance, whereas items 1-9 higher scores corresponded to better self-rated performance. Each operator's score per scenario period was the averaged ratings of the ten items.

Self-rated performance questionnaire items were selected to address multiple aspects of operator performance. The context for the questionnaire is the entire previous scenario period. Since the items address multiple aspects of an operator performance, the specificity (or grain size) for these items in self-assessment is low. Taken in its entirety, the questionnaire elicited the operator self-assessed level of performance during the past scenario-period.

4.5.4 Situation Awareness and Self-Assessment of SA

SA was assessed using the Process Overview Measure (Lau et al., 2010), which employed queries to elicit operators' knowledge on changes in plant parameters that were

relevant to the scenario events and general operating conditions (see Appendix F). Each item asked operators whether a parameter had "increased", "decreased" or "remained the same" since some cuing event in the scenario period (e.g., since a specific alarm occurred). Additionally, operators provided a confidence rating ("not confident", "neutral", or "confident") on their responses to individual queries.

Table 1 Sample Process Overview Measure cue prompt and query.

Compared to its value when the "charging pumps discharge header high-low flow" alarm (ALB 06-1-1) was received,			
1. Median Tavg Recorder indication is now:	Lower	Same	Higher
What is your confidence in your answer?	Not Conf	Neutral	Conf

During each simulator freeze, operators and the *simulator operator* (i.e., an process expert in the observation gallery operating the simulator) answered six Process Overview queries. Process Overview Measure scores were calculated by comparing the responses of the operators to those provided by the simulator operator. Operator responses that were the same as the simulator operator's response were considered correct. The final score for each participant per scenario period was the proportion of correct responses to the queries.

The confidence ratings in their responses to the queries were assigned "-1" for "not confident", "0" for "neutral", and "1" for "confident". The final score of the scenario period for each participant was the average of these confidence ratings. Since these confidence ratings were provided on operator assessments of individual parameters, they are considered to have a high level of specificity.

4.6 Hypotheses

4.6.1 Measurement Sensitivity and Validity

A portion of the literature focuses on how aspects of human performance change with different levels of difficulty. Thus, there exists a fairly large body of work on the nature of these relationships. In this experiment, scenario periods rated with *higher difficulty* indices were hypothesized to:

1. decrease task performance (both expert- and self-rated),
2. increase cognitive workload, and
3. decrease situation awareness and the corresponding confidence ratings.

4.6.2 Metacognition

Previous studies in NPP literature focusing on metacognition have demonstrated the utility of overall performance-level self-assessments in simulator studies. Thus, there reason to investigate whether operators can be specific about their performance.

Expert operators are highly trained and likely possess a deep knowledge about different aspects of their own performance. Simply stated, operators are expected to accurately self-assess performance at different levels of specificity. In particular, operators should be able to distinguish their monitoring and overall task performance and vice versa, leading to the following hypotheses:

1. Operator self-rated performance should be more indicative of actual performance than confidence assessments on specific parameter changes.

2. Operator confidence assessments on specific parameter changes (i.e., monitoring SA) should be more indicative of parameter change assessments than self-assessments on overall performance.

5 ANALYSIS & RESULTS

This chapter presents the results on the two primary objectives of this thesis in two sections. The first section examines the descriptive statistics and psychometric properties of the collected measures to evaluate their sensitivity, validity and reliability (Demas et al., 2015). After presenting the measurement properties, the second section examines the relationships between four different measures—OPAS, self-rated performance, Process Overview and Process Overview confidence—using generalized linear mixed-models to examine the specificity of operator metacognition.

5.1 Psychometric Assessment of Human Performance Measures

To provide a basic indication of effectiveness for the human performance measures, descriptive statistics of the measurements were reviewed and correlation statistics between measures were examined (Demas et al., 2015).

5.1.1 Descriptive Statistics

Table 2 presents the descriptive statistics of the human performance measures. The Wilks-Shapiro tests (Table 2, last column) indicated that the distributions of all measurements deviated significantly from normality. Hence, non-parametric statistics were employed.

Ceiling and flooring effects were inspected to assess potential issues associated with measurement sensitivity. Difficulty indices per scenario period did not show any ceiling or flooring effects. One scenario period was excluded from this analysis due to a modification of a malfunction event that occurred during the first experimental trial (resulting in n=19,

10 scenarios x 2 periods – 1 period). Thus, *a priori* difficulty indices were unavailable for that scenario period.

Table 2 Table containing descriptive statistics of human performance measures (aggregated by period).

Measure	n	Mean	S.D.	Median	Min.	Max	Skew	Kurtosis	W (p-value)
Difficulty Index	19	12.74	1.50	12.50	9.00	15.75	-0.39	0.47	0.95 (0.02)
Halden Task Complexity	180	3.55	1.00	3.40	1.20	6.80	0.50	0.43	0.97 (0.00)
OPAS	60	2.38	0.59	2.55	0.32	3.00	-1.70	2.76	0.82 (0.00)
Self-rated Performance	180	3.89	0.77	4.00	1.20	5.00	-0.91	1.10	0.94 (0.00)
Process Overview	180	0.69	0.22	0.67	0.17	1.00	-0.37	-0.69	0.92 (0.00)
Process Overview Confidence	177	0.42	0.43	0.50	-1.00	1.00	-0.78	0.13	0.93 (0.00)

The Halden Task Complexity scale scores did not show any ceiling or flooring effects (n=180, 30 scenarios x 2 periods x 3 operators). The distribution was slightly skewed-right indicating a greater proportion of low workload ratings. The Halden Task Complexity scale showed a moderate to high internal consistency (Cronbach's $\alpha = 0.78$).

The OPAS scores showed a ceiling effect (see Figure 4), indicating that task performance was not well differentiated. Twelve OPAS items were excluded from the analysis, because they were not scored at the time of data collection. However, this removal did not affect the number of scenario period scores (n = 60, 30 scenarios x 2 periods). The distribution was skewed-left, indicating a greater proportion of high performance scores. The intra-class correlation coefficients (ICCs; Shrout & Fleiss, 1979) for rater

interchangeability (ICC(2,1)) and consistency (ICC(3,1)) were both 0.85, indicating that the OPAS ratings were highly reliable across raters.

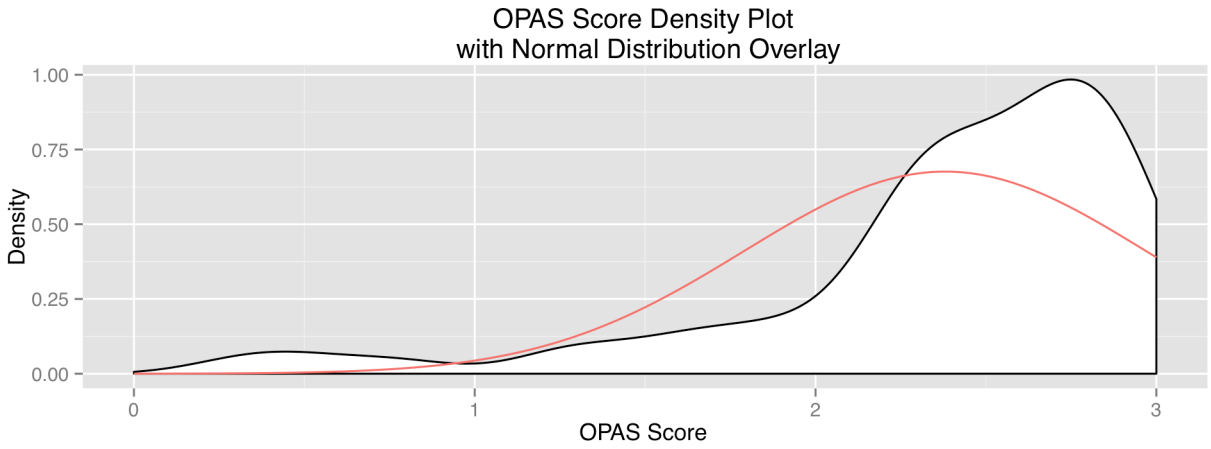

Figure 4 Density plot of OPAS scores with normal distribution overlay.

The self-rated task performance scores did not show any ceiling or flooring effects (n=180, 30 scenarios x 2 periods x 3 operators). The distribution was skewed-left, indicating a greater proportion of high self-rated performances. The self-rated task performance scores showed high internal consistency (Cronbach's α=0.95).

The Process Overview Measure distribution did not show any flooring or ceiling effects (n=180, 30 scenarios x 2 periods x 3 operators). The distribution was skewed-left, indicating a greater proportion of high scores.

The Process Overview Measure confidence scores did not show any ceiling or flooring effects. Process Overview Measure confidence scores for one participant were not answered for three periods (resulting in n=177, 30 scenarios x 2 periods x 3 operators – 3

periods). The distribution was skewed-left, indicating a greater proportion of higher scores.

5.1.2 Measurement Correlations

Non-parametric correlation statistics (Kendall's τ) were calculated to test the relationships between the human performance measurements and difficulty indices. For correlations associated with expert-rated task performance, the Halden Task Complexity scale, self-rated performance, Process Overview and Process Overview confidence scores of the three operators were aggregated into a "crew score" per scenario period (N=60).

Table 3 presents Kendall's τ, the p-values, and the number of degrees-of-freedom for the relationships between the difficulty indices and human performance measurements. Column 2 of Table 3 illustrates the relationships between the (*a priori*) difficulty indices per period and the human performance measures. The difficulty indices correlated *positively* with the Halden Task Complexity scale scores (τ=0.14, p=0.01), confirming the hypothesis that operator workload increased as scenario periods became more difficult to handle. Further, the difficulty indices correlated *negatively* with OPAS (τ=−0.19, p=0.04) and self-rated task performance (τ=−0.11, p=0.04) scores, confirming the hypothesis that operator task performance decreased as scenario periods became more difficult. Difficulty indices did not correlate with Process Overview scores, providing no support for the hypothesis on SA. However, difficulty indices correlated *negatively* with Process Overview confidence (τ=−0.12, p=0.04), confirming the hypothesis that operators were less confident in assessing their SA as scenario periods became more difficult.

The Halden Task Complexity scale also alluded to the difficulty of the scenario periods as perceived by the operators/participants. The Halden Task Complexity scale scores correlated *negatively* with OPAS ($\tau=-0.19$, $p=0.04$) and self-rated task performance ($\tau=-0.39$, $p<0.01$), conforming to the general expectation from the literature that operator task performance decreased at higher levels of mental workload. The Halden Task Complexity scale scores did not correlate with the Process Overview Measure. However, the Halden Task Complexity scale correlated *negatively* with Process Overview confidence ($\tau=-0.31$, $p<0.01$), confirming the general expectation that operators are less confident in assessing their SA at high mental workload.

OPAS ratings correlated *positively* with self-rated task performance ($\tau=0.21$, $p=0.02$), conforming to the general expectation that expert observers and the operators agreed on their performance level. OPAS did not correlate with the Process Overview Measure or Process Overview confidence.

The self-rated task performance scores correlated *positively* with Process Overview confidence ($\tau=0.21$, $p<0.01$), confirming the expectation that operator confidence was consistent between their SA and overall performance. However, self-rated task performance did not correlate with the Process Overview Measure.

Process Overview Measure scores did not correlate with Process Overview Measure confidence scores.

Table 3 Table of correlations between human performance measures.

		Halden Task Complexity	OPAS (crew level)	Self-Rated Performance	Process Overview (PO)	PO Confidence
Difficulty	**τ**	0.14	-0.19	-0.11	0.06	-0.12
	p	0.01	0.04	0.04	0.34	0.04
	df	169	55**	169	169	167**
Halden Task Complexity	**τ**		-0.19	-0.39	0.09	-0.31
	p		0.04	0	0.12	0
	df		58	178	178	175
OPAS (crew level)	**τ**			0.21	0.15	0.03
	p			0.02	0.1	0.78
	df			58	58	58
Self-Rated Performance	**τ**				-0.11	0.21
	p				0.06	0
	df				178	175
Process Overview (PO)	**τ**					-0.026
	p					0.65
	df					175

*Three difficulty scores were removed as described in Sec. 5.1.1 due to the removal of difficulty scores assessed a posteriori.

**In addition to the removal of three difficulty scores, two additional periods of confidence scores were missing from one participant.

5.1.3 Summary

Most measurements collected during this experiment displayed adequate validity and reliability. The Halden Task Complexity Scale, both measures of task performance (OPAS and self-rated performance), and Process Overview confidence responded as expected to the difficulty manipulation. Additionally, these measures displayed the expected correlations with other human performance measures. The properties displayed by the aforementioned measures demonstrate suitability for further use in simulator experiments.

5.2 Specificity of Metacognition Analysis

5.2.1 Modeling Approach

To address the hypotheses concerning the relationship between self-assessments at different levels of specificity, two sets of generalized linear mixed-models were created with self-rated performance and Process Overview Measure confidence as predictors and OPAS and Process Overview Measure scores serving as the responses.

Mixed models separate subject variation ("random effects") from the modeled quantities of interest ("fixed effects")[5]. Maximum likelihood estimation (MLE) was used to find the optimal parameters for both linear and generalized linear mixed-models. The following statistics are relevant for models obtained with MLE:

---

[5] The R package "lme4" was used to create all the generalized linear mixed-models (Bates, Machler, Bolker, & Walker, 2014; Bates et al., 2015).

1. Likelihood Ratio Test: A comparison that tests the null hypothesis that a smaller model is the same as a larger, more complicated model by comparing each model's "maximized log-likelihood [function]" (Moscatelli, Mezzetti, & Lacquaniti, 2012, p. 5). A test that fails to reject the null hypothesis indicates that smaller model is not statistically significantly different from the larger model.
2. Conditional and Marginal Adjusted R^2 (Johnson, 2014): A statistic that is interpreted as the proportion of response variable's variance explained by predictor variables. An adjusted R^2 value close to 1.0 indicates that the predictor variables explain nearly 100% of the variance in the response.

5.2.2 Data Transformations

The data collected in this experiment were aggregated to different "levels" (see Table 4) to allow for meaningful comparisons and to satisfy the condition of observation independence.

Table 4 Table presenting the level of aggregation for variables presented in different models.

Response	Predictor	Aggregation Level	# of Obs.	Missing Values
OPAS	Self-Rated Performance	Crew-Period	60	0
	Process Overview Measure Confidence	Crew-Period	60	0
Process Overview Measure Correctness	Self-Rated Performance	Subject-Period	177	3
	Process Overview Measure Confidence	Item	1058	22

Self-rated performance scores were standardized using the z-transformation (Eqn. 1) at both the crew and participant levels for each scenario period. The z-transformation was necessary to promote greater ease in interpretation of model coefficients, as the transformation maintains a score distribution's shape and centers the distribution on the mean. Standardized scores have the property that one unit change in the transformed distribution corresponds to one standard deviation in the original distribution. Equation 1 displays the z-transformation equation:

$$z_x = \frac{x - \bar{X}}{s}, \tag{1}$$

where z_x is the transformed score value, x is the untransformed score value, $\bar{X}$ is the mean of the untransformed score population and s is the standard deviation of the untransformed score population.

5.2.3 Model Variable Structure

The response variable in the models of Process Overview Measure correctness with self-rated performance as the predictor were structured as a binomial random variable of the form $\binom{n}{k}$ where n was the total number of Process Overview Measure items answered in a given scenario period by a single participant (typically 6) and k is the number of queries answered correctly during a single scenario period by a single participant.

The response for models of Process Overview Measure correctness with Process Overview Measure confidence as the predictor were treated as binary variable with "1" corresponding to a correct response and "0" corresponding to an incorrect response.

Process Overview Measure confidence was treated as a three-level categorical variable. "Neutral" was coded as the base level for the dummy-predictor variable.

5.2.4 Model Creation

Four sets of generalized linear mixed models were built to test the hypotheses on the specificity of operator metacognition. These models were generated to illustrate whether operators can differentiate between their own performance on SA queries (based on the Process Overview Measure) and overall task performance (based on OPAS and self-rated task performance). Models were created with actual task performance (OPAS scores) and SA correctness (Process Overview Measure item correctness) as responses and self-rated performance and SA confidence (Process Overview Measure item confidence) as predictors.

5.2.4.1 Mixed Model Random Effects Treatment

Two univariate linear mixed-models of OPAS scores were created for each self-assessment measure—a model with only a random intercept term and a model with both random intercept and random slope terms (see Eqn.'s 2 and 3, respectively).

$$\mathrm{y(x)} = \beta_0 + u_i^0 + \beta_1 x_{ij} \tag{2}$$

$$\mathrm{y(x)} = \beta_0 + u_i^0 + (\beta_1 + u_i^1) x_{ij} \tag{3}$$

Two univariate generalized linear mixed-models of Process Overview Measure correctness were created using the logit link function—a model with only a random intercept term and a model with both a random intercept and a random slope term (see Eqn.'s 4 and 5, respectively).

$$\log(odds) = \beta_0 + u_i^0 + \beta_1 x_{ij} \quad (4)$$

$$\log(odds) = \beta_0 + u_i^0 + (\beta_1 + u_i^1) x_{ij} \quad (5)$$

The remainder of this document presents only the results from the random-intercept models, because likelihood ratio tests between these models and the larger random-intercept and slope models did not reject the null hypothesis (i.e., conclusions drawn across the four models are practically the same).

5.2.4.2 Predicting Task Performance

Task performance was modeled with two operator self-assessment measures: (1) self-rated task performance and (2) Process Overview Measure confidence.

5.2.4.2.1 OPAS and Self-Rated Performance Model

In the model with self-rated performance as a predictor of task performance (i.e., OPAS), the intercept term β_0 represents the expert-rated performance of the crews at the average self-rated crew performance level; β_1 represents the increase in the OPAS score that results from an one-unit increase in standardized crew self-rated performance; u_i^0 represents the random-effects of different crews deviating from the β_0 intercept.

5.2.4.2.2 OPAS and Process Overview Measure Confidence Model

In the model with Process Overview Measure confidence as a predictor of task performance (i.e., OPAS), the intercept term β_0 represents the expert-rated performance when the average crew response is neutral (i.e. "0"); β_1 represents the increase in OPAS scores that results from an increase in average crew confidence from one confidence level

to the next (i.e. from "not confident" to "neutral" or "neutral" to "confident"); u_i^0 and u_i^1 represents the deviation from the β_0 intercept and β_1 slope for each crew.

5.2.4.3 Predicting Situation Awareness/Process Overview

Process Overview Measure item correctness was modeled using self-rated task performance and operator confidence on individual Process Overview Measure queries.

5.2.4.3.1 Process Overview Measure Correctness and Self-Rated Performance Model

In the model with self-rated performance as a predictor, the intercept term β_0 represents the log-odds of a correct versus incorrect Process Overview Measure response at the overall average self-rated crew performance level; β_1 represents the change in the log-odds of a correct versus incorrect response score that corresponds to a unit increase in standardized self-rated performance; u_i^0 represents the deviation from the β_0 intercept.

5.2.4.3.2 Process Overview Measure Correctness and Process Overview Measure Confidence Model

In the model of Process Overview Measure item confidence as a predictor, the intercept term β_0 represents the log-odds of an operator responding correctly versus incorrectly to a Process Overview Measure query item when the operator has neutral confidence; β_1^C represents the change in log-odds of responding correctly versus incorrectly when the operator is "Confident" versus "Neutral"; β_1^{NC} represents the change in log-odds of responding correctly versus incorrectly when the operator responds "Not Confident" versus "Neutral"; u_i^0, represents the deviation from the β_0 intercept.

5.3 Modeling Results: Specificity of Metacognition

5.3.1 Models of Expert-Rated Task Performance

The following section presents and compares the results of the two linear mixed models predicting expert-rated task performance scores with self-rated performance and Process Overview Measure confidence, respectively.

5.3.1.1 Results: OPAS and Self-Rated Performance

Table 5 contains the parameter estimates, standard error and t-values for the model predicting OPAS with self-rated performance (Eqn. 6). The marginal and conditional adjusted R^2 for the random-intercept model are 0.34 and 0.48, respectively, indicating that the fixed-effects alone account for 34% of the variance in the expert-rated task performance scores and both fixed and random-effects account for 48% of the variance in the expert-rated task performance scores.

$$y(x) = \beta_0^A + u_i^0 + \beta_1^A x_{ij} \tag{6}$$

Table 5 Estimates and confidence intervals for random intercept model of OPAS and self-rated performance.

	Estimate	Std. Error	t value	LL	UL
β_0^A	2.38	0.16	15.33	2.08	2.69
β_1^A	0.39	0.08	4.88	0.23	0.54

The fixed effect term β_1 indicates that a unit increase in crew-averaged self-rated performance corresponded to an increase in OPAS scores of 0.39.

The random-effect of crew has a standard deviation of 0.25. The normal Q-Q plot for the random effects shows that the random effects are roughly normally distributed.

Table 6 Standard deviation of random-effects for random intercept model of OPAS and self-rated performance.

Groups	Name	Std.Dev.
Crew	β_0^A	0.25
Residual		0.48

5.3.1.2 OPAS and Process Overview Measure Confidence

Table 7 contains the parameter estimates, standard error and t-values for the model predicting OPAS with Process Overview Measure confidence (Eqn. 7). The marginal and conditional adjusted R^2 are 0.13 and 0.33, respectively, indicating that the fixed effects alone account for 13% of the variation in expert-rated task performance and both fixed and random effects account for 33% of the variation.

$$\mathrm{y(x)} = \beta_0^B + u_i^0 + \beta_1^B x_{ij} \tag{7}$$

Table 7 Estimates and confidence intervals for random intercept model of OPAS and Process Overview Measure confidence.

	Estimate	Std. Error	t value	LL	UL
β_0^B	2.04	0.22	9.21	1.60	2.47
β_1^B	0.80	0.29	2.76	0.23	1.37

The random effect of crew has a standard deviation of 0.29. The normal Q-Q plot for the random effects shows that the random effects are roughly normally distributed.

The fixed effect term β_1 indicates that increases in crew-averaged Process Overview Measure confidence of one level corresponded to increases in OPAS scores of 0.80.

Table 8 Standard deviation of random-effects for random intercept model of OPAS and Process Overview Measure confidence.

Groups	Name	Std.Dev.
Crew	β_0^B	0.29
Residual		0.53

5.3.1.3 Summary of Models Predicting Task Performance

The models of OPAS scores with self-rated performance and Process Overview Measure confidence as predictors provide support for the hypothesis that operator self-assessments of overall performance are more indicative of their overall performance than self-assessments on monitoring alone. That is, self-assessment at specificity on the task level predicts task performance better than self-assessment at specificity on the monitoring SA level.

5.3.2 Process Overview Measure Correctness Models

The following section presents and compares the results of the two generalized linear mixed models predicting Process Overview Measure correctness with self-rated performance and Process Overview Measure confidence, respectively.

5.3.2.1 Results: Process Overview Measure Correctness and Self-Rated Performance

Table 9 contains the parameter estimates, standard error and z-values for the model predicting Process Overview Measure correctness with self-rated performance (Eqn. 8).

This model is not significant according to the likelihood ratio test ($\chi^2(1) = 0.077$, p=0.78) and as evidenced by the confidence intervals for the random-intercept overlapping with zero.

The lack of significant fixed-effect parameters in this model indicates that changes in self-rated performance assessment do not correspond significantly with an increase or decrease in the log-odds of correct versus incorrect Process Overview Measure answers.

$$\log(odds) = \beta_0^C + u_i^0 + \beta_1^C x_{ij} \quad (8)$$

Table 9 Estimates and confidence intervals for random intercept model of Process Overview Measure correctness and self-rated performance.

	Estimate	Std. Error	z value	Pr(>|z|)	LL	UL
β_0^C	0.81	0.14	5.96	0.00	0.54	1.07
β_1^C	-0.02	0.09	-0.28	0.78	-0.19	0.15

Table 10 Standard deviation of random-effects for random intercept model of Process Overview Measure correctness and self-rated performance.

Groups	Name	Std.Dev.
Crew	β_0^C	0.35

5.3.2.2 Results: Process Overview Measure Correctness and Process Overview Measure Confidence

Table 11 contains the parameter estimates, standard error and z-values for the model predicting Process Overview Measure correctness with Process Overview Measure

confidence (Eqn. 9). Model utility tests reveal that the model is significantly better than the null model ($\chi^2(2) = 13.813$, p=0.001).

The marginal and conditional adjusted R^2 values for the random-intercept model are 0.0212 and 0.0725, respectively, indicating that the fixed effects alone account for 2.12% of the variation in Process Overview correctness and both the fixed and random effects together account for 7.25% of the variation in Process Overview correctness (see Table 14). Though small, the effect represents a significant improvement from chance.

$$\log(odds) = \beta_0^D + u_i^0 + \beta_1^D x_{ij} \quad (9)$$

Table 11 Estimates and confidence intervals for random intercept model of Process Overview Measure correctness and Process Overview Measure confidence.

	Estimate	Std. Error	z value	Pr(>|z|)	LL	UL
β_0^D	0.54	0.19	2.87	0.00	0.17	0.90
$\beta_{1\,conf}^D$	0.52	0.16	3.24	0.00	0.20	0.83
$\beta_{1\,not\ conf}^D$	-0.11	0.24	-0.45	0.66	-0.58	0.37

The random effect of crew in the random intercept model has a standard deviation of 0.427. The normal Q-Q plot for the random effects shows that the random effects are roughly normally distributed.

The fixed-effects coefficients indicate that “Not Confident” responses are not significantly different from “Neutral” responses, but that “Confident” responses are significantly different from both “Not Confident” and “Neutral” responses.

These confidence assessments represent deviations from the overall proportion of correctness of 69% (see Eqn. 10 and Table 13). When an operator responds to a Process Overview Query and is "Confident" he has a 74% chance of answering correctly. If an operator is "Not Confident", he is equally likely to be correct or incorrect (i.e. the confidence interval for the parameter estimate includes 0.5). While not significantly different from "Not Confident" responses, the "Neutral" responses are slightly more likely to produce a correct response. Additionally, the confidence interval for this parameter estimate is above chance. Taken together, these results suggest that operators can nearly resolve these two confidence states.

$$P(correct) = \frac{e^{\log(odds)}}{1+e^{\log(odds)}} \tag{10}$$

Table 12 Standard deviation of random-effects for random intercept model of Process Overview Measure correctness and Process Overview Measure confidence.

Groups	Name	Std.Dev.
Crew	β_0^D	0.43

Table 13 Probability of correct response given different confidence levels with confidence intervals.

	Est	LL	UL
P(correct|not confident)	0.61	0.49	0.71
P(correct|neutral)	0.63	0.54	0.71
P(correct|confident)	0.74	0.67	0.80

5.3.2.3 Summary of Models Predicting SA

The model of Process Overview Measure correctness with Process Overview Measure confidence as a predictor had a higher adjusted R^2 value (i.e. explained more variance in the response variable) than the model with self-rated performance as the predictor. This result provides support to the hypothesis that operator confidence in their SA is more indicative of their SA than self-assessments of their task performance. As presented earlier, the model of expert-rated task performance with self-rated performance as a predictor had a higher adjusted R^2 (explained more variance in the response variable) than the model with Process Overview Measure confidence as a predictor. Taken together, the model results (summarized in Table 14) indicate that operators can discriminate between their task performance and their SA.

Table 14 Table of adjusted R2 values for linear and generalized linear mixed models.

	Self-Rated Performance		**POM Confidence**	
	R2m (Fixed-Effects)	*R2c (Fixed and Random Effects)*	*R2m (Fixed-Effects)*	*R2c (Fixed and Random Effects)*
OPAS	0.34	0.48	0.13	0.33
POM	0.00	0.03	0.02	0.07

5.4 Summary

The results of this experiment provide support for both hypotheses proposed in Sec. 4.6. The scenario-period difficulty manipulation produced the expected responses in four of the five human performance measures employed – the Halden Task Complexity, OPAS, self-rated performance, and Process Overview Measure confidence. Generally, these four measures also co-varied as expected and featured adequate reliability statistics. The models of both overall task performance and Process Overview Measure correctness provided support for the hypotheses on the specificity of operator metacognition.

6 DISCUSSION

This chapter begins with a discussion on the results with respect to the two research objectives: (1) psychometric evaluation of human performance measures (Demas et al., 2015) and (2) specificity of operator metacognition. The chapter continues with the implications of the research findings for ISV research and NPP modernization and safety. The chapter concludes with limitations and future work of this research.

6.1 Basic Psychometric Evaluation of Common Psychological Measures

The measurement properties of commonly used human performance measures—workload (Halden Task Complexity scale), expert-rated task performance (OPAS), self-rated task performance, and SA (the Process Overview Measure)—were examined in a full-scope simulator study (Demas et al., 2015). This paper presented results on the basic measurement properties of and interrelationships between these four common human performance measures.

The experimental results support the use of the Halden Task Complexity scale to measure workload in full-scope simulator experiments and ISV activities. The Halden Task Complexity scale measurements increased with *a priori* difficulty indices of scenario periods, demonstrating basic validity and sensitivity. Further, the Halden Task Complexity scale scores also correlated negatively with expert- and self- rated task performance, and SA confidence in the expected direction. This measure also showed adequate internal consistency (Braarud, 2001; Skraaning Jr. et al., 2007). The results from this full-scope simulator experiment supplement the empirical evidence on the Halden Task Complexity scale (Braarud, 2000, 2001).

The preliminary results generally support the use of OPAS to measure task performance. OPAS scores decreased with increasing *a priori* difficulty indices, demonstrating basic validity and sensitivity. Additionally, OPAS scores correlated negatively with the Halden Task Complexity scale and positively with self-rated performance as expected. OPAS also had high inter-rater reliability, corroborating previous findings (Lau, Jamieson, & Skraaning Jr, 2012). However, visual inspection of OPAS measurements revealed a ceiling effect, indicating poor differentiation of performances for this study. Human factors researchers and process experts must devote attention to item criteria for better resolution of task performance in future experiments.

The results support the use of the self-rated performance questionnaire. Self-rated performance questionnaire scores decreased with increasing *a priori* difficulty indices, demonstrating basic validity and sensitivity. Self-rated performance scores also correlated negatively with the Halden Task Complexity scale and positively with both OPAS and SA confidence scores, as expected. The self-rated performance questionnaire displayed very high internal consistency, indicating that all items measured the same psychological construct. Though corroborating with internal consistency results in Skraaning Jr. et al. (2007), Cronbach's α above 0.90 raised the concern that some items appeared redundant. Future investigation may consider eliminating or replacing several items.

The experimental results do not provide strong support for the use of the Process Overview Measure. The measure did not respond to scenario difficulty or correlate with OPAS or the Halden Task Complexity scale. The lack of correlation with OPAS and Halden Task Complexity corroborated with previous findings (Lau et al., 2010; O'Brien & O'Hare, 2007). Thus, the SA results must be interpreted carefully (Lau & Skraaning Jr., 2015).

However, the results indicate the merits of the Process Overview confidence ratings. Process Overview confidence decreased with increasing *a priori* difficulty and cognitive workload. Additionally, Process Overview confidence was high when self-rated performance was also high.

6.2 Metacognitive Specificity of Task Performance and SA

The second objective of this thesis was to explore and provide evidence for the specificity of operator self-assessments about different aspects of their performance. To address this objective, four mixed-models were constructed to determine if operators could distinguish between self-assessments of task performance and SA. The results of these models (1) suggest that operator self-assessments of overall task performance (i.e. lower specificity self-assessments) better predict actual task performance than self-assessments of SA and (2) suggest that operator self-assessments on individual SA queries (i.e. high-specificity self-assessments) better predict SA than self-assessments given on their overall performance.

The model of actual task performance with self-rated performance suggests that operators can provide self-assessments that are indicative of their actual performance when asked about their overall performance (low specificity) better than when asked about their SA (higher specificity). The model of Process Overview Measure correctness with Process Overview Measure confidence assessments as a predictor suggest that when asked about their own SA (higher specificity), operators can provide self-assessments that are more indicative of their SA than self-assessments on their general level of performance (lower specificity). Taken together, these results indicate that when operators provide

confidence assessments on their SA they are able to distinguish that these assessments are different from their assessments of their overall performance.

The crew-averaged SA confidence assessments were also indicative of actual task performance, but to a lesser degree than self-rated performance. The self-rated performance questionnaire addressed multiple aspects of operator performance including aspects of monitoring. Process Overview Measure confidence represents self-assessments on monitoring obtained during the evaluation of specific parameter changes that were then averaged. Thus, it can be argued that the Process Overview Measure confidence scores represent self-assessment on the monitoring aspect of performance.

However, the lack of a significant relationship between self-rated performance and Process Overview Measure correctness provides some insight into the relationship between overall performance and SA. One might expect that overall self-assessments of performance represent a base level of confidence and that when asked questions of a more specific nature, corrections would be made to that baseline. In agreement with the existing literature (O'Brien & O'Hare, 2007), the non-result (between Process Overview Measure correctness and self-rated performance) can be viewed as a form of discriminate validity.

It appears that operators are able to be more specific with their self-assessments *only when asked*. That is, self-assessments resulting from being asked about monitoring *in general* (i.e. through the self-rated performance questionnaire) are different from self-assessments of confidence on specific aspects of monitoring (i.e. from Process Overview Measure Confidence). This finding on both high and low specificity self-assessments absent

in the literature demonstrates that operators are capable of distinguishing their overall task performance from their SA performance.

6.3 Practical Implications

The psychometric evaluations of multiple measures provide several benefits for both ISV practitioners and regulators. Practitioners now have access to information about how these measures behave. In addition, the measures have been translated from Swedish to American NPP operator English. For regulators, the CAER facility has demonstrated its capabilities for conducting full-scope NPP control room experiments and can be used for research to inform future guidance.

Current NRC regulations require assessing plant performance, task performance, workload, and situation awareness during ISV activities. Despite the demonstrated importance of metacognition in nuclear power and other domains, the regulations currently do not require measurement of the construct. This absence may be due to limited number of studies on the construct targeting the nuclear domain in the literature. From a practical standpoint, this thesis demonstrates that operator metacognition on two different aspects of performance can be assessed through minimal addition to measures currently required in regulation.

This work will be especially helpful to regulators given the impending modernization and new plant builds that must evaluate automated safety systems to ensure they promote safe plant operation. These automated systems are designed to provide a high degree of safety. However, this safety is contingent upon appropriate use of automation. As discussed in Sec. 2.4.5, the proper usage of automation depends on an

operator's mental comparison between his own ability and his perception of the automation's capabilities. If an operator trusts his own ability less that the automation's capability, he is likely to delegate more tasks to the automation regardless of whether this delegation is appropriate. This thesis addressed the component of this comparison involving an operator's self-assessment and showed that operators can be specific about their self-assessments. This work parallel's investigations into the specificity of operator assessments of automated system functions (Lee & Moray, 1992).

Operators must work in conjunction with automation in process control settings. In doing so, operators must be aware of both the process parameters under control and the state and functioning of the automation. By utilizing the Process Overview Measure, this thesis addressed the specificity of operator awareness of their own process monitoring. Future evaluations could apply the results presented in this thesis on the level of specificity self-assessment to determine better methods of automated system assessments in control room evaluations.

6.4 Limitations and Future Work

This thesis demonstrates that operator assessments of overall task-performance are more indicative of actual task performance than SA monitoring confidence, and that operator assessments of SA monitoring confidence are more indicative of SA monitoring correctness. Thus, while it demonstrates that operators *can be* more specific, it does not provide a measure answering *how* specific they can be.

The Process Overview Measure has been established as a sensitive a valid indicator of human performance in nuclear process control. However, the measure did not garner full

support in this experiment, as it did not respond to the difficulty manipulation nor did it co-vary as expected with any other of the human performance measures. Even though the exact nature of the relationship between SA and other measures is contentious in the literature, the lack of support complicates the interpretation of the results.

This experiment compared self-assessments on overall task performance with performance on SA queries. Operator specificity on other human performance constructs was not explored. Future experiments should examine other constructs that include trust in automation, perceived plant performance, and automated system performance.

6.5 Conclusion

This thesis addressed both practical and research needs relevant to the nuclear industry, including human performance measure evaluation and exploration. The thesis provides psychometric support of domain specific measures of human performance as well as support for the inclusion of metacognition in NPP evaluation. In doing so, this thesis contributes to more comprehensive evaluations of control room technologies that in turn may be implemented in real world settings thereby providing a reasonable assurance of safe operation and the potential for increased levels of plant productivity.

7 CONCLUSION

Nuclear power provides a viable option to achieve emissions goals set forth by the US EPA. However, utilities must demonstrate safe plant operation through testing of control room technologies with expert operators using scientifically sound, multidimensional tests to provide the public the greatest assurance of safety. This thesis presented the results of a full-scope NPP simulator experiment that evaluated the basic psychometric properties of multiple human performance measures and provided evidence on a novel human performance construct—metacognition. Through these results the psychometric properties of several human performance measures currently required by NUREG-0711 were collectively validated for further use in control room assessments. Additionally, models of operator performance and SA demonstrated that operators can be specific in their self-assessments concerning different aspects of their performance. These results have implications for modernization and new construction projects for both ISV practitioners and regulators.

REFERENCES

Baranski, J. V, & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics*, *55*(4), 412–428. doi:10.3758/BF03205299

Baron, L. (2014). Personal Communication.

Barriere, M., Bley, D., Cooper, S., Forester, J., Kolaczkowski, A., Luckas, W., … Whitehead, D. (2000). *Technical basis and implementation guidelines for a technique for human event analysis (ATHEANA)*. *NUREG-1624, Rev* (Vol. 1).

Boring, R., Lew, R., Ulrich, T., & Joe, J. (2014). *Light Water Reactor Sustainability Program Operator Performance Metrics for Control Room Modernization: A Practical Guide for Early Design Evaluation (INL/EXT-14-31511)*. Idaho National Laboratory (INL).

Braarud, P. Ø. (2000). *Subjective task complexity in the control room (HWR-621)*. Institutt for energiteknikk, OECD Halden Reactor Project, Halden (Norway).

Braarud, P. Ø. (2001). Subjective task complexity and subjective workload: Criterion validity for complex team tasks. *International Journal of Cognitive Ergonomics*, *5*(3), 261–273.

Braarud, P. Ø., & Skraaning Jr., G. (2006). Insights from a benchmark integrated system validation of a modernized NPP control room: performance measurement and the comparison to the benchmark system. In *Proceedings of the 5. International Topical Meeting on Nuclear Plant Instrumentation Controls, and Human Machine Interface Technology*.

Braarud, P. Ø., & Svengren, H. (2006). Impact of an Ambiguous Secondary Task on Primary Task Performance in Accident Operation. *PSAM8, New Orleans, USA*.

Broberg, H., Hildebrandt, M., Massaiu, S., & Braarud, P. Ø. (2008). A Simulator Experiment Investigating the Effects of Masked Indicators in a NPP Emergency Control Task: Scenario and Design. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 52, pp. 1728–1732).

Burns, C. M., Skraaning, G., Jamieson, G. A., Lau, N., Kwok, J., Welch, R., & Andresen, G. (2008). Evaluation of ecological interface design for nuclear process control: situation awareness effects. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(4), 663–679.

Bye, A., Lois, E., Dang, V. N., Parry, G., Forester, J., Massaiu, S., … Julius, J. (2011). *International HRA Empirical Study–Phase 2 Report: Results from Comparing HRA*

Method Predictions to Simulator Data from SGTR Scenarios (NUREG/IA-0216, Vol. 2). US Nuclear Regulatory Commission.

Cain, B. (2007). *A Review of the Mental Workload literature (RTO-TR-HFM-121-Part-II)*. Defense Research and Development Canada Toronto.

Carvalho, P. V. R., dos Santos, I. L., & Vidal, M. C. R. (2006). Safety implications of cultural and cognitive issues in nuclear power plant operation. *Applied Ergonomics*, *37*(2), 211–223. doi:10.1016/j.apergo.2005.03.004

Charlton, S. G., & O'Brien, T. G. (2001). *Handbook of human factors testing and evaluation*. CRC Press.

Chuang, C.-F., & Chou, H.-P. (2008). Design development and implementation of the human-system interface for Lungmen nuclear project. *Nuclear Science, IEEE Transactions on*, *55*(5), 2654. doi:10.1109/TNS.2008.2003977

Conca, J. (2014, July 1). Are EPA's Carbon Rules Really About Nuclear? *Forbes*.

Cooper, G. E., & Harper Jr, R. P. (1969). *The Use of Pilot Rating in the Evaluation of Aircraft Handling Qualities (No. AGARD-567)*. ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT NEUILLY-SUR-SEINE (FRANCE).

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. doi:10.1007/BF02310555

De Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies*, *58*(6), 719–735. doi:http://dx.doi.org/10.1016/S1071-5819(03)00039-9

Demas, M. W., Lau, N., & Elks, C. E. (2015). Advancing Human Performance Assessment Capabilities for Integrated System Validation -- A Human-in-the-Loop Experiment. In *Proceedings of the 9th International Conference on Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies (NPIC and HMIT 2015)*.

Dong, X., Song, F., Li, Z., & Zhang, S. (2013). Data extraction and analysis for integrated system validation of a nuclear power plant. *Nuclear Engineering and Design*, *265*, 826–832.

Droeivoldsmo, A., Skraaning Jr., G., Sverrbo, M., Dalen, J., Grimstad, T., & Andresen, G. (1998). *Continuous Measures of Situation Awareness and Workload (HWR-539)*. OECD Halden Reactor Project, Halden, Norway.

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and

accessibility hypotheses. *Journal of Memory and Language*, *52*(4), 565. doi://dx.doi.org/10.1016/j.jml.2005.01.011

Edgar, G. K., Edgar, H. E., & Curry, M. B. (2003). Using Signal Detection Theory to Measure Situation Awareness in Command and Control. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *47*(18), 2019–2023. doi:10.1177/154193120304701815

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). In *Aerospace and Electronics Conference, 1988. NAECON 1988., Proceedings of the IEEE 1988 National* (pp. 789–795). doi:10.1109/NAECON.1988.195097

Endsley, M. R. (1995a). Measurement of Situation Awareness in Dynamic Systems, *37*(1), 65–84.

Endsley, M. R. (1995b). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 32.

Endsley, M. R. (2015). Final Reflections: Situation Awareness Models and Measures. *Journal of Cognitive Engineering and Decision Making*, *9*(1), 101–111. doi:10.1177/1555343415573911

Fader, G. (2009). New Nuclear Plant Deployment. IEEE Meeting April 2009 -- INPO New Plant Deployment. Retrieved from http://ewh.ieee.org/r3/atlanta/ias/IEEE_April_2009.ppt

Flach, J. M. (1995). Situation Awareness: Proceed with Caution. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 149–157. doi:10.1518/001872095779049480

Flach, J. M., Mulder, M., & van Paassen, M. M. (2004). The concept of the situation in psychology. *A Cognitive Approach to Situation Awareness: Theory and Application*, 42–60.

Gawron, V. J. (2008). *Human performance, workload, and situational awareness measures handbook*. CRC Press.

Goldstein, S. (1986, August 22). Soviets Say Six Human Errors Led To Chernobyl. *Philly.com*. Philadelphia. Retrieved from http://articles.philly.com/1986-08-22/news/26063183_1_chernobyl-reactor-valery-legasov-soviet-officials

Guznov, S., Reinerman-Jones, L., & Marble, J. (2012). Applicability of Situation Awareness and Workload Metrics for Use in Assessing Nuclear Power Plant Designs. In *Advances in Cognitive Engineering and Neuroergonomics*.

Ha, J. S., & Seong, P. H. (2009). HUPESS: Human performance evaluation support system. In *Reliability and Risk Issues in Large Scale Safety-critical Digital Control Systems* (pp. 197–229). Springer.

Ha, J. S., Seong, P. H., Lee, M. S., & Hong, J. H. (2007). Development of human performance measures for human factors validation in the advanced MCR of APR-1400. *Nuclear Science, IEEE Transactions on*, *54*(6), 2687–2700.

Hallbert, B. P. (1997). Situation awareness and operator performance: results from simulator-based studies. *Proceedings of the 1997 IEEE Sixth Conference on Human Factors and Power Plants, 1997. "Global Perspectives of Human Factors in Power Generation."* doi:10.1109/HFPP.1997.624933

Hamblin, C. J., Castaneda, M., Fuld, R. B., Holden, K., Whitmore, M., & Wilkinson, C. (2013). Verification and Validation Human Factors Requirements and Performance Evaluation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, *52*, 139–183.

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner, A. C., Zaklade, A. L., & Christ, R. E. (1992). Comparison of Four Subjective Workload Rating Scales. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *34*(4), 429–439.

Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust . *Human Factors: The Journal of the Human Factors and Ergonomics Society* , *57* (3), 407–434. doi:10.1177/0018720814547570

Hogg, D. N., Folleso, K., Strandvolden, F., & Torralba, B. (1995). Development of a Situation Awareness Measure to Evaluate Advanced Alarm Systems in Nuclear-Power-Plant Control Rooms. *Ergonomics*. doi:10.1080/00140139508925275

Hwang, S.-L., Yau, Y.-J., Lin, Y.-T., Chen, J.-H., Huang, T.-H., Yenn, T.-C., & Hsu, C.-C. (2008). Predicting work performance in nuclear power plants. *Safety Science*, *46*(7), 1115–1124. doi:10.1016/j.ssci.2007.06.005

INPO. (2014). INPO - About Us. Retrieved July 14, 2014, from http://www.inpo.info/AboutUs.htm

Jang, I., Park, J., & Seong, P. (2012). An empirical study on the relationships between functional performance measure and task performance measure in NPP MCR. *Annals of Nuclear Energy*, *42*, 96–103. doi:10.1016/j.anucene.2011.10.004

Jeannot, E. (2000). *Situation Awareness: Synthesis of Literature Search (EEC Note No. 16/00)*. European Organisation for the Safety of Air Navigation, Eurocontrol Experimental Centre.

Kato, M. (2014). Mythbusting with Manga of Fukushima Cleanup. *Wired*. Retrieved July 15, 2014, from http://www.wired.co.uk/magazine/archive/2014/07/play/mythbusting-manga

Kim, A. R., Lee, S. W., Park, J., Kang, H. G., & Seong, P. H. (2013). Correlation analysis between team communication characteristics and frequency of inappropriate communications. *Annals of Nuclear Energy*, *58*, 80–89. doi:10.1016/j.anucene.2013.03.003

Kim, A. R., Park, J., Lee, S. W., Jang, I., Kang, H. G., & Seong, P. H. (2012). Development of new taxonomy of inappropriate communication and its application to operating teams in nuclear power plants. *Nuclear Engineering and Technology*, *44*(8), 897–910. doi:10.5516/NET.04.2011.068

Kim, J. T., Shin, S. K., Kim, J. H., & Seong, P. H. (2013). An experimental approach to estimate operator's information processing capacity for diagnosing tasks in NPPs. *Annals of Nuclear Energy*, *59*, 100–110. doi:10.1016/j.anucene.2013.03.023

Krug, K. (2007). The relationship between confidence and accuracy: Current thoughts of the literature and a new area of research. *Applied Psychology in Criminal Justice*, *3*(1), 7.

Lang, A. W., Roth, E. M., Bladh, K., & Hine, R. (2002). Using a Benchmark-Referenced Approach for Validating a Power Plant Control Room: Results of the Baseline Study. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 46, pp. 1878–1882). doi:10.1177/154193120204602302

Lau, N., Jamieson, G. A., & Skraaning Jr, G. (2012). Inter-rater Reliability of Expert-Based Performance Measures. In *Proceedings of the 8th American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation & Control and Human-Machine Interface Technologies (NPIC & HMIT), San Diego, CA, USA* (pp. 1974–1982).

Lau, N., Jamieson, G. A., Skraaning Jr., G., & Burns, C. M. (2008). Providing Operator Support during Monitoring for Unanticipated Events through Ecological Interface Design. *Proc. of the 29th Annual Canadian Nuclear Society Conference*.

Lau, N., Skraaning, G., Jamieson, G. A., & Burns, C. M. (2008). Enhancing Operator Task Performance During Monitoring for Unanticipated Events Through Ecological Interface Design. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 52, pp. 448–452).

Lau, N., Skraaning Jr, G., & Jamieson, G. A. (2009). Metacognition in Nuclear Process Control. In *Proc. of the 17th Triennial World Congress on Ergonomics [CD-Rom]*.

Lau, N., & Skraaning Jr., G. (2015). Exploring sub-dimensions of situation awareness to support integrated system validation. In *Proceedings of the 9th International*

Conference on Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies (NPIC and HMIT 2015). Charlotte, NC, USA.

Lau, N., Skraaning Jr., G., Eitrheim, M. H., Karlsson, T., Nihlwing, C., & Jamieson, G. A. (2010). *Situation awareness in monitoring nuclear power plants: The Process Overview concept and measure (HWR-954)*. OECD Halden Reactor Project, Halden, Norway.

Le Blanc, K. L., Boring, R. L., & Gertman, D. I. (2001). *Review of methods related to assessing human performance in nuclear power plant control room simulations (INL/CON-10-19525)*. Idaho National Laboratory (INL).

Lee, J. D. (1999). Measuring Driver Adaptation to In-Vehicle Information Systems: Disassociation of Subjective and Objective Situation Awareness Measures. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. doi:10.1177/154193129904301810

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153–184. doi:http://dx.doi.org/10.1006/ijhc.1994.1007

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15151155

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270.

Leis, R., Reinerman-Jones, L., Mercado, J., Barber, D., & Sollins, B. (2014). Workload from Nuclear Power Plant Task Types Across Repeated Sessions. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 58, pp. 210–214).

Lichacz, F. M. J. (2008). Augmenting understanding of the relationship between situation awareness and confidence using calibration analysis. *Ergonomics*, *51*(10), 1489–1502. doi:10.1080/00140130802277513

Lichacz, F. M. J., Cain, B., & Patel, S. (2003). Calibration of Confidence in Situation Awareness Queries. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *47*(1), 222–226. doi:10.1177/154193120304700147

Lichacz, F. M. J., & Farrell, P. S. E. (2005). The calibration of situation awareness and confidence within a multinational operational net assessment. *Military Psychology*, *17*(4), 247–268. doi:10.1207/s15327876mp1704_1

Lichtenstein, S., Fischhoff, B., & Phillips, L. (1982). Calibration of probabilities: The state of the art to 1980. D. Kahneman, P. Slovic, and A. Tverski (Eds.) Judgement under uncertainty: Heuristics and biases. New York, Cambridge University Press.

Lin, C. J., Hsieh, T. L., Tsai, P. J., Yang, C. W., & Yenn, T. C. (2011). Development of a team workload assessment technique for the main control room of advanced nuclear power plants. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *21*(4), 397–411. doi:10.1002/hfm.20247

Lysaght, R. J., Hill, S. G., Dick, A. O., Plamondon, B. D., & Linton, P. M. (1989). *Operator workload: Comprehensive review and evaluation of operator workload methodologies (TR-2075-3)*. ANALYTICS INC WILLOW GROVE PA.

Macfarlane, A. M. (2014). Investing in Safety: The Importance of Effective Regulation. *NRC News*. Retrieved from http://www.nrc.gov/reading-rm/doc-collections/commission/speeches/2014/s-14-001.pdf

Magill, B. (2015, January 29). Nuclear Power Needs to Double to Meet Warming Goal | Climate Central. *Climate Central*. Retrieved from http://www.climatecentral.org/news/nuclear-power-needs-to-double-to-meet-warming-goal-18610

Matthews, G., Reinerman-Jones, L. E., Barber, D. J., & Abich, J. (2015). The Psychometrics of Mental Workload: Multiple Measures Are Sensitive but Divergent . *Human Factors: The Journal of the Human Factors and Ergonomics Society* , *57* (1), 125–143. doi:10.1177/0018720814539505

McGuinness, B. (2004). *Quantitative analysis of situational awareness (QUASA): Applying signal detection theory to true/false probes and self-ratings*. *2004 Command and Control Research and Technology Symposium*.

Mercado, J. E., Reinerman-Jones, L., Barber, D., & Leis, R. (2014). Investigating Workload Measures in the Nuclear Domain. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 58, pp. 205–209).

Meshkati, N. (1991). Human factors in large-scale technological systems' accidents: Three Mile Island, Bhopal, Chernobyl. *Organization & Environment*, *5*(2), 133.

Miberg Skjerve, A. B., & Bye, A. (2011). *Simulator-based Human Factors Studies Across 25 Years: The History of the Halden Man-Machine Laboratory*. Springer.

Mitsubishi Heavy Industries Ltd. (2008). *US-APWR Human System Interface Verification and Validation (Phase 1a)*. MUAP-08014-NP (R0).

Moracho, M. J. (1998). Plant Performance Assessment System (PPAS) for crew performance evaluation. Lessons learned from an alarm system study conducted in HAMMLAB (HWR-504). Institutt for energiteknikk, OECD Halden Reactor Project, Halden (Norway).

Moscatelli, A., Mezzetti, M., & Lacquaniti, F. (2012). Modeling psychophysical data at the population-level: The generalized linear mixed model. *Journal of Vision*, *12*(11). doi:10.1167/12.11.26

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429–460. doi:10.1080/00140139608964474

Murphy, K. R., & Davidshofer, C. O. (2001). *Psychological testing: principles and applications* (6th ed.). Prentice Hall.

Norros, L., & Nuutinen, M. (2005). Performance-based usability evaluation of a safety information and alarm system. *International Journal of Human-Computer Studies*, *63*(3), 328.

Nuclear-Energy-Institute. (2015a). License Renewal of Nuclear Energy Facilities. Retrieved from http://www.nei.org/Master-Document-Folder/Backgrounders/Fact-Sheets/Relicensing-Nuclear-Energy-Facilities?feed=factsheet

Nuclear-Energy-Institute. (2015b). New Nuclear Energy Facilities Will Support Growth, Provide Clean Electricity. Retrieved from http://www.nei.org/Master-Document-Folder/Backgrounders/Fact-Sheets/New-Nuclear-Energy-Facilities-Will-Support-Growth,

O'Brien, K., & O'Hare, D. (2007). Situational awareness ability and cognitive skills training in a complex real-world task. *Ergonomics*, *50*(7), 1064–1091.

O'Hara, J. M., Higgins, J. C., Fleger, S. A., & Pieringer, P. A. (2012). Human Factors Engineering Program Review Model (NUREG-0711, Revision 3). United States Nuclear Regulatory Commission.

O'Hara, J. M., Stubler, W., Higgins, J., & Brown, W. (1997). *Integrated system validation: methodology and review criteria*. Division of Reactor Controls and Human Factors, Office of Nuclear Reactor Regulation, US Nuclear Regulatory Commission.

Plumer, B. (2014, June 1). A guide to Obama's new rules to cut carbon emissions from power plants. *Vox*. Retrieved from http://www.vox.com/2014/6/1/5770556/EPA-power-plant-rules-explainer

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. *Advances in Psychology*, *52*, 185–218.

Reinerman-Jones, L., Guznov, S., Mercado, J., & D'Agostino, A. (2013). Developing Methodology for Experimentation Using a Nuclear Power Plant Simulator (pp. 181–188). Springer.

Roberts, R., Flin, R., & Cleland, J. (2014). Staying in the Zone: Offshore Drillers? Situation Awareness. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(4), 573–590. doi:10.1177/0018720814562643

Roth, E. M., Easter, J., Hall, R. E., Kabana, L., Mashio, K., Hanada, S., … Remley, G. W. (2010). Person-in-the-Loop Testing of a Digital Power Plant Control Room. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

Rueb, J., Vidulich, M., & Hassoun, J. (1992). Establishing workload acceptability: An evaluation of a proposed KC-135 cockpit redesign. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

Sarter, N. B., & Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology*, *1*(1), 45–47.

Schoenherr, J., Leth-Steensen, C., & Petrusic, W. (2010). Selective attention and subjective confidence calibration. *Attention, Perception, & Psychophysics*, *72*(2), 353–368. Retrieved from http://dx.doi.org/10.3758/APP.72.2.353

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science*, *26*(1), 113–125. doi:10.1023/A:1003044231033

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, *4*(1), 33–45.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Skjerve, A. B., Nihlwing, C., Nystad, E., & Strand, S. (2009). Lessons Learned from the Extended Teamwork Study on Work Organization, Collaboration Technologies, and Automatic Agents. In *Proceedings of 17th World Congress on Ergonomics*. Bejing, China.

Skjerve, A. B., & Strand, S. (2007). Teamwork in a new operational environment in nuclear power plants. Preliminary results from the extended teamwork study. In *Proceedings of the workshop on future control station designs and human performance issues in nuclear power plants*. Halden, Norway.

Skraaning Jr, G., Lau, N., Welch, R., Nihlwing, C., Brevig, L. H., Veland, Ø., … Kwok, J. (2007). *The Ecological Interface Design Experiment (2005, HWR-833)*. OECD Halden Reactor Project, Halden (Norway).

Skraaning Jr., G. (1998). *The operator performance assessment system (OPAS, HWR-538)*. Halden, Norway: OECD Halden Reactor Project.

Skraaning Jr., G., Eitrheim, M. H. R., & Lau, N. (2010). Coping with Automation in Future Plants. In *Seventh American Nuclear Society International Topical Meeting on Nuclear*

Plant Instrumentation, Control and Human-Machine Interface Technologies (pp. 1–9). Las Vegas, NV.

Skraaning Jr., G., & Miberg Skjerve, A. B. (2006). Trust in automation and meta-cognitive accuracy in NPP operating crews. In *Proceedings of the 5th International Topical Meeting on Nuclear Plant Instrumentation Controls, and Human Machine Interface Technology*.

Stanton, N. A., Salmon, P. M., & Walker, G. H. (2015). Let the Reader Decide: A Paradigm Shift for Situation Awareness in Sociotechnical Systems . *Journal of Cognitive Engineering and Decision Making* , *9* (1), 44–50. doi:10.1177/1555343414552297

Strawn, J. J. (2010). *Assessing mental workload and situation awareness in the evaluation of real-time, critical user interfaces*. Oregon State University.

Sulistyawati, K., Wickens, C. D., & Chui, Y. P. (2011). Prediction in Situation Awareness: Confidence Bias and Underlying Cognitive Abilities. *The International Journal of Aviation Psychology*, *21*(2), 153–174. doi:10.1080/10508414.2011.556492

Swaton, E., Neboyan, V., & Lederman, L. (1987). Human factors in the operation of nuclear power plants. *IAEA Bulletin*, *29*(4), 27.

Taylor, R. M. (1990). Situational Awareness Rating Technique(SART): The development of a tool for aircrew systems design. *AGARD, Situational Awareness in Aerospace Operations 17 p(SEE N 90-28972 23-53)*.

Taylor, R. M. (1995). CC-SArt: The Development of an Experiential Measure of Cognitive Compatibility in System Design. *Report to TTCP UTP-7 Human Factors in Aircraft Environments, Annual Meeting*.

Taylor, R. M., & Selcon, S. J. (1991). Subjective measurement of situational awareness. *Designing for Everyone, Proceedings of the 11th Congress of the International Ergonomics Association (Paris)*.

Tsang, P. S., & Vidulich, M. A. (1989). Cognitive demands of automation in aviation. In R. S. Jensen (Ed.), *Aviation psychology*. Brookfield, VT: Gower Publishing Co.

Ulrich, T., Boring, R., Phoenix, W., Dehority, E., Whiting, T., Morrell, J., & Backstrom, R. (2012). *Applying Human Factors Evaluation and Design Guidance to a Nuclear Power Plant Digital Control System (INL/EXT-12-26787)*. Idaho National Laboratory (INL).

Van Winsen, R., & Dekker, S. W. A. (2015). SA Anno 1995: A Commitment to the 17th Century . *Journal of Cognitive Engineering and Decision Making* , *9* (1), 51–54. doi:10.1177/1555343414557035

Vicente, K. J., Mumaw, R. J., & Roth, E. M. (2004). Operator monitoring in a complex dynamic work environment: a qualitative cognitive model based on field observations. *Theoretical Issues in Ergonomics Science*, *5*(5), 359–384. doi:10.1080/14639220412331298929

Vidulich, M. A., & Hughes, E. R. (1991). Testing a subjective metric of situation awareness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

Vidulich, M. A., Stratton, M., Crabtree, M., & Wilson, G. (1994). Performance-based and physiological measures of situational awareness. *Aviation, Space, and Environmental Medicine*, *65*(5).

Waag, W. L., & Houck, M. R. (1994). Tools for assessing situational awareness in an operational fighter environment. *Aviation, Space, and Environmental Medicine*, *65*(5).

Walker, J. S. (2000). *Short History of Nuclear Regulation, 1946-1999*. DIANE Publishing.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 449.

Wickens, C. D. (2015). Situation Awareness: Its Applications Value and Its Fuzzy Dichotomies . *Journal of Cognitive Engineering and Decision Making* , *9* (1), 90–94. doi:10.1177/1555343414564571

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). *Engineering Psychology and Human Performance*. Prentice Hall New Jersey.

Wierwille, W. W., & Casali, J. G. (1983). A validated rating scale for global mental workload measurement applications. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

Wierwille, W. W., & Eggemeier, F. T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *35*(2), 263–281.

Wilson, G. F. (2000). Strategies for psychophysiological assessment of situation awareness. In M. R. Endsley & D. J. Garland (Eds.), *Situation Awareness Analysis and Measurement* (p. 175). CRC Press.

World-Nuclear-Association. (2015). Nuclear Power in the USA. Retrieved from http://www.world-nuclear.org/info/Country-Profiles/Countries-T-Z/USA--Nuclear-Power/

APPENDIX A

Background Questionnaire

1. Age: ____yrs

2. Gender:
    a) Male ____
    b) Female ____

3. How long have you retired from being an nuclear power plant operator: ____yrs

4. Which nuclear power plant did you last from: _____

5. Role in this Experiment:
    a) Reactor operator: ☐
    b) Senior reactor operator: ☐
    c) Unit supervisor: ☐

6. Describe your formal education

7. Which licenses did you last hold?
    a) Reactor operator: ☐
    b) Senior reactor operator: ☐
    c) Unit supervisor: ☐

8. How many years have you worked in the following positions:
    a) Reactor operator: _____yrs
    b) Senior reactor operator: _____yrs
    c) Unit supervisor: _____yrs

9. How many years have you worked in the following positions of nuclear power plants
    a) Electrical maintenance: _____yrs
    b) Instrument maintenance: _____yrs

c) Mechanical maintenance: _____yrs
d) Instructor: _____yrs
e) Field operator: _____yrs
f) Shift Engineer: _____yrs
g) Other: _____yrs

10. What is your main/current position in your nuclear power plant? (Checked all that applies.)
    a) Reactor operator: ☐
    b) Senior reactor operator: ☐
    c) Unit supervisor: ☐
    d) Electrical maintenance: ☐
    e) Instrument maintenance: ☐
    f) Mechanical maintenance: ☐
    g) Instructor: ☐
    h) Field operator: ☐
    i) Shift engineer: ☐
    j) Other: ☐

Computer experience

11. How frequently do you use computers at work?
    a) Several hours each day ☐
    b) A few hours each week ☐
    c) A few hours each month ☐

12. Do you use a computerized control room panels/interfaces in connection with operation (i.e., monitoring and control) of the plant?
    a) Never ☐
    b) For a few tasks ☐
    c) For most tasks ☐
    d) For all tasks ☐

13. If you have used a computerized control room panels/interfaces during operation, for how many years: _____yrs

14. In general, what do you think about using computerized control room panels in operation of the plant?

APPENDIX B

Semi-Structured Post-Experiment Interview

Overall Experience

1. How do you feel about the overall experience operating through the scenarios in this experiment?
2. Would you be interested in participating in such experiment again?

Control Room Setup

1. The control room setup is built entirely of computer monitors that are quite different from most nuclear power plant control to date. Do you think the setup is adequate for research purposes? For instance, studying the usefulness of the large screen displays? [This facility is not meant to induce/replicate exact operator behaviors in real control rooms, only an approximation.]
2. What do you think of a "Windows" type of displays containing graphics as in the large screen displays? Do you think you can operate the simulator based on such interfaces? [E.g., the AP1000 or petrochemical plant displays.]
3. What additional tools would you like to see being added in this control room that could really support you operating through the scenarios?

Crew/Participant Composition

1. This experiment only includes Senior Reactor Operator, Reactor Operator, and Unit Supervisor role. The Shift Manager and other roles are not there. Do you think the composition is adequate for research purposes (e.g., studying large screen displays)?
2. What about an even smaller crew for preliminary testing of displays or concepts? Do you think a crew of Senior Reactor Operator and a Unit Supervisor can manage to operate through some scenarios?

Large Screen Display

1. What do you think about the large screen displays? Useful or not useful? Explain.
2. Any likes and dislikes? What improvements would you like to see?
3. Would you like to see some of those features on the desktop monitors?

Human Performance Measures

1. What do you think of the pauses during the scenarios for answering various questionnaires? [How did they affect your workflow?]
2. Do you have any comments for the questionnaires:

    a. Complexity rating
    b. Self-rating
    c. Process parameter queries
    d. Confidence rating
3. How do you feel about wearing the eye-tracking glasses? Do they affect your workflow?
4. How do you feel about wearing the breathing-rate sensor belt? Do they affect your workflow?

Training & Scenario Design

1. Given that all of you are not licensed on this particular simulator, do you feel the training provided to you is "adequate" for operating through the scenarios?
2. How can we improve the training for someone without experience for this plant model?
3. How can we improve the scenario design for someone without experience for this plant model?

Schedule

1. We have completed 10 runs for the entire experiment within one week. What do you think about this "workload"?

Future Topics

1. We are interested in running experiments related to technology for modernization and new constructions for the future. What do you think of the following topics:
    a. Digital I&C and automation failures
    b. Computerized procedures
    c. Cyber response management

What else do you think we need to consider for improving this facility for research to support the nuclear industry?

APPENDIX C

Modified Halden Task Complexity Scale

Table III contains the modified version of the Halden Task Complexity scale (Braarud, 2000, 2001) utilized in this experiment. Item wording was modified to closer reflect American NPP operator English, operative phrases were bolded for greater clarity, and item phrasing was altered to create self-contained statements.

Table 15 Modified Halden Task Complexity scale questionnaire used in this experiment.

1. I found the information on displays ambiguous, misleading or missing.
2. I found feedback from my control actions ambiguous, misleading or missing.
3. I found time a factor for planning and responding to the plant event/disturbance.
4. I found executing every single task complicated by many simultaneous tasks (several disturbances or plant events).
5. I found collecting information to handle the plant disturbance to be difficult.

APPENDIX D

Sample OPAS Items

Table 16 contains a sample OPAS (Skraaning Jr., 1998) rating sheet used in the experiment.

Table 16 Sample OPAS rating items used in this experiment.

Event: Selected Feed Flow channel (FT-497) fails high. 'C' SGWLC responds by reducing FF to 'C' SG.		
Expected Operator Actions	**Range of Performance**	
DETECT • FI-497 increases (selected 'C' FF failing high) • 'C' MFRV demand decreases • Actual FF to 'C' SG decreases (FI-496) • 'C' SG level decreases. • ALB 14-3-1B ('C' Level error) received at 52%	Expert use of available diverse indications.	3
	Minor delays in checking diverse indications.	2
	Significant lapses in use of available indications delay response to failure.	1
	Diverse indications not used effectively.	0
DIAGNOSE • Observes conflicting 'C' FF indications • Observes decreasing 'C' SG level • Observes decreasing demand on 'C' MFRV • Determines selected channel of 'C' FF is failing high	Timely & logical diagnosis using diverse indications.	3
	Minor lapses in diagnostics but recovers.	2
	Significant delays diagnosing failure result in 'C' SG level less than 40% NR.	1
	Misinterpreted indications lead to incorrect diagnosis (opposite to existing failure).	0
RESPOND • Determines MAN control of 'C' MFRV is required. • Places 'C' MFRV in MAN • Raises FF to recover 'C' SG level to 57% • Adjusts 'C' FF to stabilize 'C' SG level at program	Response is timely and precise. Level controlled in near program with minimal overshoot.	3
	Some delay in response. Some overshoot occurs.	2
	Significant delays. "Heavy-handed" response results in wide variations of 'C' SG level and feed flow.	1
	No response before Rx Trip. Response occurs but in wrong direction (without recovery). 'C' SG level decreases below 30%. Significant lapses in monitoring.	0
COORDINATE: • Performs ALB 14-3-1B and OWP-RP • Deselects Ch-3 FF (& SF) • Restores 'C' MFRV to AUTO • Monitors for proper operation of 'C' MFRV • Initiates corrective actions & makes notifications (I&C and OMOC)	Performed correctly. Smooth transition MAN to AUTO control w/o Level Error alarm. Notifications made.	3
	Minor lapses in performance. Level error alarm occurs.	2
	Significant lapses. SF not deselected. Some notifications made.	1
	'C' MFRV remains in MAN (no attempt to deselect failed channel and return to AUTO). No notifications.	0

APPENDIX E

Modified Self-Rated Task Performance Questionnaire

Table 17 shows the modified version of the self-rated performance questionnaire (Skraaning Jr. et al., 2007) used in this experiment. Items 1, 2, 3, 6, 7 of the original questionnaire were modified to reflect language used by American NPP operators. Items 8 and 10 of this questionnaire were added to address other aspects of operator performance.

Table 17 Modified self-rated performance questionnaire used in this experiment.

1. I maintained a good overview of the plant conditions.
2. I carried out my actions in a timely manner.
3. I communicated well with the crew.
4. I made correct diagnoses.
5. My actions within the team steered the response in the correct direction.
6. I utilized the displayed alarms, indications, and controls effectively.
7. I became aware of pertinent changes in plant conditions at an early stage.
8. I felt that I fulfilled my responsibilities in my shift position.
9. I performed the correct control actions.
10. Sometimes during the scenario, I did not understand the plant conditions.

APPENDIX F

Sample Process Overview Measure Items

Table 18 contains a sample Process Overview Measure (Lau et al., 2010) query used in this experiment.

Table 18 Sample Process Overview Measure query used in this experiment.

Compared to its value when the "charging pumps discharge header high-low flow" alarm (ALB 06-1-1) was received,			
1. Median Tavg Recorder indication is now:	Lower	Same	Higher
What is your confidence in your answer?	Not Conf	Neutral	Conf
2. Pressurizer Level indication (LI-461) is now:	Lower	Same	Higher
What is your confidence in your answer?	Not Conf	Neutral	Conf
3. Main Generator Gross Electrical Output is now:	Lower	Same	Higher
What is your confidence in your answer?	Not Conf	Neutral	Conf
4. VCT Level indication (LI-115) is now:	Lower	Same	Higher
What is your confidence in your answer?	Not Conf	Neutral	Conf
5. Charging Flow indication (FI-122) is now:	Lower	Same	Higher
What is your confidence in your answer?	Not Conf	Neutral	Conf
6. RHX Letdown Temperature indication (TI-140) is now:	Lower	Same	Higher
What is your confidence in your answer?	Not Conf	Neutral	Conf