AN ABSTRACT OF THE THESIS OF

Jordan J. Strawn for the degree of Master of Science in Computer Science presented on June 4, 2010.

Title: Assessing Mental Workload and Situation Awareness in the Evaluation of Real-Time, Critical User Interfaces.

Abstract approved:

_____

Carlos Jensen

While there are many ways to evaluate a user interface design, the user's mental workload and situation awareness (SA) are particularly important considerations in the supervisory control of safety-critical systems. Typically, operators of these systems must monitor high-volume, time-sensitive status information. Interface design for this domain can be challenging and should consider both workload and SA, because presenting too much information can lead to cognitive overload. For highly automated systems, operator underload and poor SA are also potential issues. This study reviewed subjective and objective techniques for assessing workload and SA, and analyzed the suitability of each. In recommending a suite of measures, we sought general applicability to the real-time, critical application domain, but we focused particularly on next-generation multi-modular reactor control rooms. Based on a human factors experiment in which each participant monitored and controlled multiple simulated reactors, we recommend the NASA-TLX instrument and a chest-band heart rate monitor for assessing mental workload. For the

assessment of SA, we recommend using eye gaze data. In support of summarizing the results of user interface evaluation along multiple dimensions (e.g., workload, SA, user error), we propose that these inter-related constructs be presented in a single, standardized, task-based format.

Assessing Mental Workload and Situation Awareness in the Evaluation of Real-Time, Critical User Interfaces

by

Jordan J. Strawn

A THESIS

submitted to

Oregon State University

in partial fulfillment of

the requirements for the

degree of

Master of Science

Presented June 4, 2010

Commencement June 2010

Master of Science thesis of Jordan J. Strawn presented on June 4, 2010.

APPROVED:

_____

Major Professor, representing Computer Science

_____

Director of the School of Electrical Engineering and Computer Science

_____

Dean of the Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

_____

Jordan J. Strawn, Author

ACKNOWLEDGEMENTS

I am extremely grateful to the many people who contributed their time and insights in support of this research effort. First of all, thank you to Carlos, my advisor, for all of the guidance along the way, for your willingness and flexibility to let my research follow my interests, for your legwork to obtain funding for various possible paths, and for your support of the lab setup efforts. I enjoyed our conversations about sports, particularly your experiences at the Olympics. I also thank the rest of my committee for taking interest in this work and giving their time and helpful feedback: Drs. Margaret Burnett and Ron Metoyer, Computer Science; Dr. Mei-Ching Lien, Psychology; Dr. Michael Scott, Structural Engineering (Graduate Council Representative).

Thank you to the OSU HCI research group for providing advice on the experimental protocol and particularly to Jose Cedeno, Jennifer Davidson and Nitin Mohan for participating in our pre-experiment pilot. Karthick Subramanian and Vignesh Viswanathan made substantial contributions to the literature review, the experimental preparations and administration, and data analysis. I owe them a huge thanks and wish them great success in their future research efforts.

NuScale Power supported this work, both financially and with the valuable time and insights of its employees. Thank you to Ken Harris, manager of Instrumentation and Controls, for giving me the opportunity to intern with NuScale over the summer of 2009. Ken, I enjoyed working under you and diving into control room human factors and NUREGs, even though my desk was in the kitchen. Steven Blomgren, Brandon Haugh and Ross Snuggerud supported our efforts as subject matter experts. Dr. Jose Reyes, Jr.

obtained sample operating procedures for us from Trojan Nuclear Power Plant and established contacts for us in the OSU Nuclear Engineering Department. Jose, I hold you in very high personal and professional esteem, and I'm grateful for your support and interest in my academic progress over the years, despite your hectic schedule.

Dr. Ken Funk of the OSU Dept. of Industrial Engineering and Heather Lonsdale, a graduate of our research group, provided early insights on human factors evaluation from their respective domains of aviation and UAV control interfaces. From the OSU Dept. of Nuclear Engineering, Dr. Brian Woods provided us with a reactor simulator from the IAEA and assisted with recruitment. Dr. Qiao Wu kindly gave us a tour of the APEX test facility and its control area. On short notice, Robert Schickler and Gary Wachs showed us OSU's TRIGA reactor control room and answered our human factors questions.

We are grateful to have received in-person demonstrations and expert advice on a number of technical topics, which proved indispensable to our understanding of the domain and our analysis efforts. Dr. Mei-Ching Lien, Alison Gemperle and Nathan Herdener of the OSU Dept. of Psychology's Attention and Performance Lab kindly demonstrated EEG data collection and analysis. Jay Penry of the OSU Dept. of Nutrition and Exercise Sciences showed us EKG, provided very helpful advice and pointed us to useful papers. Dr. Anthony Hornof, Yunfeng Zhang and Kyle Vessey of the Computer and Information Science Department at the University of Oregon demonstrated two eye tracking systems to us and shared practical advice from their lab experience with the technology.

Todd Shechter, Glen Winters and Justin Spencer of OSU Engineering Support were very patient and helpful in assisting us with the complex and unusual lab setup for the validation experiment. I am also grateful to the five participants who lent their time and expertise in supporting our experiment.

On a personal note, thanks to great friends, caring church members and family, for continual support, relaxing breaks from school, and for periodically asking about my research, even if you just wanted the short version. My parents, Gregg and Pam, encouraged me to pursue grad school, and have contributed to my life in countless more important ways over the years, particularly in exemplifying lives of faith, integrity and love. I am so grateful to my wife, Kellie, for her support throughout the master's program, particularly during the final months of the thesis effort. Thank you for listening to my frustrations, being patient and encouraging during the busy stretches, and for sharing wisdom and practical advice from your grad school days at Oklahoma State. We are finally both done!

As this degree appears to be the culmination of my formal education, it seems appropriate to reflect on my public education and express thanks to my teachers, instructors and professors over the years, from preschool through grad school. Incredibly, only six of the 102 individuals I was able to count as having taught me were in classrooms outside Corvallis; this minority was from a study abroad program in Vienna, Austria. Even more unlikely, the campuses of the CHS preschool, Garfield Elementary School, Highland View Middle School (now demolished), Corvallis High School (now re-built), and Oregon State University were or are all adjacent to 11th. Street in Corvallis.

It is probably a rare feat to learn so much on one street. Having spent twenty years in classrooms, I have surely learned and forgotten a lot. I received a well-rounded education including band, German and the social sciences, in addition to solid teaching in math, science and technology. I am grateful to all the teachers who were passionate about their subjects and about contributing to the lives of their students. They each have claim to at least a small, albeit anonymous, part in this accomplishment.

On a related note, I would like to acknowledge by name the memory of Dr. Curtis Cook, who taught my introductory Computer Science course at OSU and served as my undergraduate advisor until his retirement. When we first met to discuss my five-year plan, he laughed and said, “Wow, you sound like a busy beaver!” and then proceeded to describe the term in the computation sense. Dr. Cook was the first full-time faculty member in the department. He passed away this year.

Proverbs 3:5-6; Isaiah 40:28-31

J.S.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

Page

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

LIST OF TABLES

Assessing Mental Workload and Situation Awareness in the Evaluation of Real-Time, Critical User Interfaces

1. INTRODUCTION

On October 21, 2009, a commercial flight from San Diego to Minneapolis overshot its destination by 150 miles. A flight attendant on Northwest Flight 188 eventually alerted the pilot and co-pilot to their error. Fourteen minutes after the scheduled arrival time, the crew finally contacted air traffic control from the air to report their mistake, and proceeded to turn the plane around for a tardy descent into Minneapolis. Although none of the 149 passengers or crew was harmed in the incident, the White House was alerted and National Guard jets had been prepared for takeoff to intercept the wayward Airbus A320, as the pilots had neglected customary communication with controllers on the ground for 91 minutes. This bizarre incident made national headlines and the FAA revoked the licenses of both pilots for their recklessness and error. (Ahlers, 2009; Associated Press, 2009; Kavanagh, 2009)

Although experts theorized that the pilots were fatigued and had fallen asleep at the controls, the pilots explained their inattention to both instrumentation and radio contact by stating that they had been using their laptops to discuss new scheduling software. In this case, the crew clearly exhibited poor situation awareness, as they neither knew their location nor the time, despite the repeated attempts of controllers and other pilots to contact them since the plane had passed over Denver. At the time, this incident

was discussed as an especially extreme example of a more wide-spread issue in aviation: due to the high levels of automation and the auto-pilot capabilities of contemporary airplanes, pilots find it challenging to maintain attention on their very important task (Kavanagh, 2009).

While the dangers of overloaded human operators in critical applications are self-evident and have long been discussed in the literature, there is a growing body of research suggesting that underload conditions may also be undesirable (e.g., Pattyn et al., 2008; Rubio et al., 2004). Therefore, while highly automated systems may be desirable for multiple reasons, human factors experts and user interface designers should pay special attention to the influence of this design decision on the workload and situation awareness of the operating crew. As exemplified in the Flight 188 incident, operators of highly automated systems may become bored and understimulated to the point of distraction and low vigilance, or alternatively, they must exert considerable effort to focus their attention in order to maintain awareness.

While the pilots involved were harshly rebuked by the FAA and their licenses were revoked, the impact of the pilot error in terms of passenger safety turned out to be minimal. Unfortunately, other instances of inattention, poor situation awareness and over-reliance on automation in aviation have contributed to fatal accidents (Endsley, 1996; Endsley et al., 2003; Kavanagh, 2009).

While much of the work in human factors has its origins in aviation, this is certainly not the only safety-critical domain in which the complex interactions of tasks, procedures, human operators, the environment and the human-system interface must be

carefully considered. In the nuclear power domain, the U.S. Nuclear Regulatory Commission (NRC) has issued multiple NUREGs (NRC regulatory guidance documents) regarding human factors in power plant design and operation. Regulatory oversight of this domain, as well as public distrust, grew considerably following the partial core meltdown accident at Three Mile Island, Unit 2 in 1979 (NRC, 2009). Fortunately, the radiation dosages in the surrounding population were minor and no lives were lost (NRC, 2009). Nevertheless, a number of NUREGs address this accident in particular, and federal law was changed based on investigations of the events ("Contents of applications; technical information,"10 CFR 50.34 (f)).

The TMI-2 accident has been explained as a combination of system malfunctions, problematic control room interface design, and human error (Endsley, 1996; NRC, 2009). Following an initial system failure, a faulty indicator design led the operators to an incorrect understanding of the developing conditions, and thus low situation awareness regarding the system state (Ha et al., 2007; NRC, 2009). At the same time, the crew was overwhelmed by a host of competing alarms, with the more instructive alarms lost in the noise. Both of these contributing factors, one related to situation awareness and the other to cognitive overload, can be traced back to the design of the control room interface. "We set them up for failure," said Dr. Jose Reyes, Jr., a member of a grand jury task force assigned to investigate the accident (Learn, 2010).

The nuclear power industry now better recognizes "human performance as a critical part of plant safety" (NRC, 2009, p. 3). For example, in response to the human-system interface design issues uncovered in the investigations into the TMI-2 accident,

U.S. federal law now requires “a control room design that reflects state-of-the-art human factor principles” (10 CFR 50.34 (f)(2)(iii)).

In developing or improving interfaces, designers must apply lessons learned from past operating incidents, and carefully consider the interaction of tasks, procedures, human strengths and limitations, and the user interface. There are various ways to evaluate a candidate interface design, including expert inspection and observation of test user interactions with the system. NUREG-0711, “Human Factors Engineering Program Review Model,” emphasizes the latter (O’Hara et al., 2004). Of particular relevance here, and motivating this research effort, NUREG-0711 calls for assessments of situation awareness and mental workload which “reflect the current state-of-the-art” (p. 70), as part of the interface design and evaluation process. However, this NUREG stops short of specifying the method of assessment; in practice, there is a diverse multitude of techniques, including subjective, physiological and performance-based measures.

The human element of a system is a timely and growing consideration, not limited to academia and safety-critical human-system interfaces. The January-February 2010 issue of *FAA Aviation News* (___, 2010) focused on human factors considerations, including pilot workload, fatigue, and human aspects of avionics certification. Also, Campbell Soup Co. recently completed a multi-year effort of evaluating consumer responses to the graphical presentations on can labels (Brat, 2010). Campbell contracted Innerscope Research Inc. to perform the evaluations consisting of a suite of physiological techniques, including eye tracking and heart rate recording, to measure the emotional

responses of participants to test labels. The results were sufficiently conclusive to initiate an overhaul of the traditional design.

Measuring human subjective and physiological responses to informational interfaces can be useful in a variety of domains. As guided by NUREG-0711, the present effort considers assessment of mental workload and situation awareness in particular. However, choosing from the vast set of potential methods is non-trivial, and should be informed by an understanding of the particulars of the system and the operator population under investigation.

This study focuses especially on the workload- and situation-awareness-based evaluation of user interfaces (also called human-system interfaces) for "real-time, critical applications." This characterization implies two things. First, information describing the system, potentially of high volume, diversity and complexity, must be presented to the user (i.e., the operator) in a timely fashion. Regarding the operators in various critical applications, Endsley stated "all must perceive and comprehend a dazzling array of data that is often changing very rapidly" (2000, p. 3). The user needs this information in order to understand the system state and to respond appropriately. Control actions may also be time-sensitive. Second, for critical applications, the safety and well-being of personnel or the general public may depend on proper operator decision-making and actions. Furthermore, operators in such critical situations are potentially subject to psychological stress, particularly when under pressure to perform correctly and efficiently. This stress is related to, but not exactly equivalent to, mental workload. Hornof et al. (2010) described this target domain as an important one in the field of Human-Computer Interaction, with

interface design requiring an understanding of human perception and cognition, especially if user multi-tasking is anticipated.

In the context of driving, Healey and Picard (2005) discussed "real-time noncritical applications" (p. 164), such as navigation aids, cell phone conversations and the car stereo. These secondary devices themselves are not safety-related, but interactions with such devices require driver attention and thus have workload and awareness implications. Another example of a real-time, non-critical domain is sports commentating, in which a commentator must make timely, interesting, correct remarks based on his or her view of the action and on statistical data available through IT systems at the venue (Midy et al., 2007). Although lives are not at stake, the individual's livelihood depends on what he or she says. It is therefore desirable for the information system to provide both an overview and noteworthy details, while avoiding information overload. Even in such non-critical domains, this can be a difficult design problem.

While intended for general applicability, the present research effort was funded by and is geared toward answering questions relevant to NuScale Power's Human Factors Engineering program. NuScale Power (http://nuscalepower.com) is a Corvallis, Oregon-based nuclear engineering and plant design firm spun out from technology developed by Oregon State University and Idaho National Environment and Engineering Laboratory (INEEL). NuScale Power is one of a number of companies seeking to develop and commercialize smaller nuclear reactors for localized or modularized power production. Extensive human factors efforts will be required during the design, validation and certification of control rooms for such next generation, or "advanced," power plants

(e.g., Tran et al., 2007b), as significant modifications are expected from the traditional horseshoe-shaped control room with analog controls and alarm annunciators.

In a traditional plant, as specified in current federal regulations, an operating crew is devoted to a single reactor unit, and at most two units are controlled from a single control room (see "Conditions of licenses," 10 CFR 50.54(m)). NuScale Power is proposing to assign each operator to monitoring multiple small, highly-automated reactor modules. This development, foreseen by Tran et al. (2007b), necessitates an exemption from the federal regulations for control room staffing. While there is an exemption process in place for such circumstances, as outlined in NUREG-1791 (Persensky et al., 2005), the application of the process is, to our knowledge, unprecedented.

In seeking exemption, NuScale Power must demonstrate that with the proposed design, "public health and safety will be maintained at a level that is comparable to compliance with the current regulations" (p. I-1-1, Persensky et al., 2005). This demonstration must include data from a number of significant activities carried out for any new design, as outlined in NUREG-0711 (O'Hara et al., 2004). As mentioned above, this document calls for state-of-the-art assessment of mental workload and situation awareness as part of an iterative human-system interface design and evaluation process. This form of experimental or simulator-based evaluation is natural when considering the allocation of functions between operators and system automation, the design of tasks and procedures, crew member coordination, and the adequacy of the user interface in supporting safe operation. All of these considerations are inter-twined with the staffing model and may impact mental workload and situation awareness. Therefore, an

understanding of the law, the plant design as it relates to control room human factors, and the NuScale concept of operations all serve as inputs to this research effort.

This thesis is organized as follows: Chapter 2 summarizes an extensive literature review. The problem constraints specific to NuScale Power's needs are examined more thoroughly in Chapter 3. Criteria for the selection of experimental measures are chosen in Chapter 4, followed by the analysis and selection process itself (Chapter 5). Chapters 6, 7 and 8 present the protocol, results and discussion, respectively, of a pilot experiment designed for validating the recommended measures. The report closes with both NuScale Power-specific and general conclusions (Chapter 9).

The goals of this research are to:

1. Select the most appropriate methods for measuring mental workload (MW) and situation awareness (SA) for the experimental evaluation of a real-time, critical user interface.
    a. Seek a general approach applicable to various related domains.
    b. In cases of competing constraints, tune the approach for the anticipated environment and tasks of a multi-modular reactor control room.
2. Apply the proposed methods to an example user interface, serving as an initial pilot study / simple validation.
3. Develop a framework / rubric for evaluating task complexity and user interface support of operator tasks, including, potentially, mental workload, situation awareness, error rates, error impact, and operator skill requirements. Such a framework may help to:

a. Determine the adequacy of a user interface for supporting safe system operation.
b. Identify problematic tasks/procedures/UI screens for improvement.

2. RELATED WORK

2.1 Concepts and Terminology in the Evaluation of User Interfaces

2.1.1 Performance

A common way to evaluate a user interface is to measure the performance of the user during interactions with the system (e.g., O'Hara et al., 2004). The system designer or user interface evaluator may choose performance measures that indicate the level of attainment of system goals. Benchmarks for satisfactory performance may be specified prior to the evaluation, based on system specifications or comparison with the human performance in an existing system (Ha et al., 2007).

In general, good user performance, meaning efficient and satisfactory completion of tasks, suggests that the user interface supports system goals adequately. Somewhat complicating the matter, there is commonly a trade-off between speed and accuracy (Proctor & Van Zandt, 2008), with task instructions influencing the relative priorities of the participant. Response time and task completion time are common measures of speed, while error rates reflect accuracy (see, e.g., Dixon & Wickens, 2003; NUREG-1791 section 10.1.2, Persensky et al., 2005).

2.1.2 Error Rates and Impact

The types of errors which users tend to commit when interacting with a user interface are important considerations during design and evaluation; also relevant are the limitations of the human as an information processor and the user's mental model (Norman, 1983). Norman presented design guidelines for minimizing the occurrence and

impact of some types of user errors. In a lab user study, Maxion and Reeder (2005) compared two user interfaces for editing file permissions based on the user errors observed. Although users spent a comparable amount of time verifying their work with each interface and subjective confidence in the task accuracy was comparable between the two, users completed the assigned tasks more quickly and with a lower incidence of goal errors with the prototyped interface. Maxion and Reeder classified the errors observed with the technique THEA (Pocock et al., 2001, in Maxion & Reeder), a framework intended for evaluation of error with user interfaces.

Because the operator is a somewhat unreliable component in a safety-critical system, analyses of human reliability and human error are important aspects of probabilistic risk assessment (PRA; see, e.g., NUREG-0711, O'Hara et al., 2004). There are many techniques in addition to THEA for analyzing human errors, such as ATHEANA, but an in-depth investigation of such techniques is out of the scope of this effort.

Of particular relevance to this research, mode errors are a situation awareness-related challenge in the supervisory control of highly automated systems. These errors describe the user's failure to understand what the automated system is doing, or providing incorrect input to the system based on the current mode. Such errors may arise due to system complexity, inaccurate mental models, the attentional demands of system monitoring, or inadequate feedback of system state and actions. Regarding feedback, the presentation of information matters, not just the volume of data provided. (Sarter & Woods, 1995)

Also of relevance, Wilson and Funk (1998) investigated task management errors in advanced cockpits by analyzing operational incident reports. Inappropriate task prioritization can potentially have devastating consequences. Having found significantly more such errors in advanced than traditional cockpits, they speculated that this difference may be due to the challenges of interacting with automated systems.

2.1.3 Mental Workload

Workload refers to the resource expenditure required of a person in performing one or more tasks (see e.g., Hart, 2006; O'Donnell & Eggemeier, 1986; Proctor & Van Zandt, 2008). Often, workload is a dynamic concept, as task demands change over time and operators adjust their strategies accordingly (Veltman & Gaillard, 1996). According to Nachreiner (1995), operator workload should be one criterion in the evaluation of system designs. Mental workload, also commonly referred to as "cognitive workload," considers perceptual and cognitive demands in particular (Plott et al., 2004), excluding other factors such as physical workload (Hwang et al., 2008). This construct matters because the human operator is viewed as having limited attentional and processing resources with which to perform tasks (Baldwin, 2003; O'Donnell & Eggemeier, 1986; Svensson et al., 1997). Mental workload is an emerging and important concept in the field of Human-Computer Interaction (Gevins & Smith, 2003; Iqbal et al., 2005), and a common human factors consideration in evaluating operator tasks and interactions with a system. However, there is no universally accepted definition (Hwang et al., 2008; Theureau, 2000; Veltman & Gaillard, 1996; Yeh & Wickens, 1988). Similarly, many

diverse methods have been proposed and used for assessing workload (O'Donnell & Eggemeier, 1986).

Proctor and Van Zandt (2008) defined mental workload as "the amount of mental work or effort necessary to perform a task in a given period of time" (p. 248). Naturally, the number and complexity of assigned tasks generally affects mental workload (O'Donnell & Eggemeier, 1986). Based on the above definition, time demands also influence workload levels. That is, if the same task must be performed with varying time constraints, the conditions with tighter deadlines could be expected to produce higher workload levels. Rowe et al. (1998) made a distinction between mental demands and mental effort expended, as an operator may direct efforts at only a subset of the total task demands; mental effort is also related to motivation. Veltman and Gaillard (1996) emphasized that the capacity of the individual is a central factor in the determination of mental workload. As an example, Smith-Jackson and Klein (2008) found differences in subjective workload ratings between individuals related to task absorption.

Stress has been described alternatively as a component of, or an effect of, workload (Gaillard, 1993). Springing from different research areas, these two concepts both concern resources and demands, and neither is well-defined (Gaillard, 1993). Healey and Picard (2005) defined stress as "a reaction from a calm state to an excited state for the purpose of preserving the integrity of the organism" (p. 156). Similarly, Gaillard (1993) described stress as "a state in which the operator feels threatened and is afraid of losing control of the situation" (p. 1002). Warm et al. (1996) suggested a relationship between task-induced stress and workload. Wientjes (1992) lumped mental effort and

"psychological stress" together in terms of their effects on respiration. Similarly, Tharion et al. (2009) investigated the physiological effects of anxiety-induced (mental) stress, rather than workload, per se.

The SWAT assessment technique incorporates stress as one of the three dimensions of subjective workload (Reid & Nygren, 1988). Vidulich and Wickens (1986) also included "Stress level" as one of eight subjective rating scales in comparing subjective and performance measures. In contrast, other studies have considered stress and fatigue as results of high workload levels, rather than as corollaries or components of workload (Healey & Picard, 2005; Tran et al., 2007a; Tran et al., 2007b). Zhang and Luximon (2005) explicitly omitted situation-induced stress and emotional factors from their formulation of mental workload, limiting it to "the demand on the brain and the sensory system (eyes, ears, and skin) due to the task" (p. 200).

Gaillard (1993) proposed a two-dimensional model describing mental effort and stress as two types of energy mobilization in response to demands on the operator. Temporarily, an intentional increase in mental effort can benefit performance. On the other hand, a sense of a lack of control may contribute to an involuntary stress response and ultimately hinder performance. Therefore, sustained workload at high levels may eventually contribute to stress (Gaillard, 1993).

A useful perspective on mental workload considers the amount of reserve capacity available for meeting additional perceptual or cognitive demands (Proctor & Van Zandt, 2008). Depletion of reserve capacity is thus synonymous with high mental workload and a decrement in performance (see Yeh & Wickens, 1988). Similarly, Wilke

et al. (1985) presented a "four-stage stress cycle," in which perceived stress stems from perceived demands exceeding available resources. With the reserve capacity view of workload and stress, overall mental workload can be decomposed into the demands placed on subsystem resources (see, e.g., Wickens, 2008). Mental workload is therefore commonly deemed a multi-dimensional construct (Baldwin, 2003; Sirevaag et al., 1993; Yeh & Wickens, 1988). In practice, operator capacity may be difficult to determine, as it is influenced by the individual's abilities, training, previous experience and present state of activation (Gaillard, 1993).

Due to the development of increasingly complex systems and the ability to display large volumes of dynamic system information, excessive mental workload, or overload, has become a concern (see, e.g., Berka et al., 2005; Brookings et al., 1996; Gevins & Smith, 2003; Hwang et al., 2008; Rowe et al., 1998; Svensson et al., 1997; Tremoulet et al., 2009). Although measures of performance and mental workload do not show a direct relationship, overload may result in decreased user performance. When stress or task demands are too great, operators may resort to "task shedding," in which they change their strategies at the expense of some task goals (Hancock & Szalma, 2003; Rowe et al., 1998). Overloaded users may also be more prone to errors (Alexander et al., 2000) and take longer to complete tasks (Plott et al., 2004). High workload levels may lead crew members to neglect proper communication and coordination, which are essential for maintaining team situation awareness (Patrick et al., 2006). O'Donnell and Eggemeier (1986) presented a hypothetical model relating performance to mental workload, in which performance can be maintained satisfactorily up to a workload cut-off

point (i.e., the transition point between regions A and B in their model); performance subsequently drops off drastically for higher workload levels (see also Yeh & Wickens, 1988). This workload threshold, defining the overload region, is called the "redline." This threshold may vary across individuals, based on relative abilities and previous experience (Proctor & Van Zandt, 2008; Gevins & Smith, 2003).

In a contrasting theoretical model relating performance to workload, conditions of excessively low workload are also undesirable (Baldwin, 2003; Braby et al., 1993; Hwang et al., 2008; Mouloua et al., 2001; Nachreiner, 1995; Pattyn et al., 2008; Proctor & Van Zandt, 2008). According to this model, performance follows an "inverted U-shaped" curve, and is optimized when users are engaged but not overwhelmed (Proctor & Van Zandt, 2008). The Yerkes-Dodson Law relates performance to arousal level along a similar inverted U (Yerkes & Dodson, 1908, in Proctor & Van Zandt, 2008). Braby et al. (1993) related underload to boredom, which is associated with both low and high arousal; their results supported the former. Also, Wilke et al. (1985) presented evidence of an inverted U function of stress in university faculty: those experiencing moderate stress levels reported higher productivity than their colleagues under either low or high stress.

Nachreiner (1995) characterized both over- and underload situations as "dysfunctional." In conditions of underload, the lack of external stimulation or task challenge may potentially result in complacency and inattention (Hwang et al., 2008); low alertness (Proctor & Van Zandt, 2008); monotony and large response times (Baldwin, 2003); boredom and errors (Tsang & Wilson, 1997, in Alexander et al., 2000). International Standard ISO 10075-2 (1996) presents ergonomic design principles related

to mental workload, including avoidance of underload in conditions of potential monotony or sustained vigilance. As indicated by the two conflicting theories relating performance to workload, underload is a complex issue: Durso et al. (1999) found that "low mental demand sometimes suggests good performance and sometimes poor performance" (p. 7).

Supervisory control is an increasingly common role for the operators of modern, highly automated systems (Gaillard, 1993; Jou et al., 2009; Tran et al., 2007a). This concept of operations features monitoring tasks with a very low event rate (i.e., less than 1-2 response-worthy events per hour; Gaillard, 1993), which is frequently assumed to induce low mental workload (Hwang et al., 2008; Jou et al., 2009; Warm et al., 1996). However, Warm et al. (1996) considered this a "myth," finding that monitoring tasks can produce fairly high workload levels (see also Tran et al., 2007a, 2007b). Baldwin (2003) characterized such vigilance tasks as "effortful and stressful" (p. 135). Furthermore, Berka et al. (2004) equated mental workload with alertness, on a "high vigilance" to "sleepy" continuum. Pattyn et al. (2008) came to somewhat different conclusions: although subjective ratings of mental workload may be high under such vigilance conditions, psychophysiological measures suggest that underload is the culprit. In effect, understimulation can lead to boredom and low activation levels, such that the human must expend mental effort in order to maintain vigilance, and thus system awareness (Gaillard, 1993). A performance decrement in vigilance tasks may be observed within twenty minutes of task initiation (Mouloua et al., 2001). While greater automation may reduce overall workload, it potentially reduces physical load at the expense of increased

mental workload (Ha et al., 2007; see also Jou et al., 2009, for a review of automation levels and mental workload). With automated process control, mental workload may rise dramatically following a disturbance, as the operator attempts to understand the system state and respond accordingly (Hallbert, 1997).

2.1.4 Situation Awareness

Situation awareness is a concept that originated in the fighter aircraft domain in the past century (Endsley, 2000). In early formulations, the term referred to a pilot's need to integrate information from multiple sources in order to achieve mission goals (Sarter & Woods, 1991). Situation awareness has since become a design consideration in many domains (Endsley, 2000), including nuclear power plant operation, in part because of the human operator's role as monitor in increasingly automated systems (Theureau, 2000). That is, automation is a closely related topic, and potential hindrance, to situation awareness (Endsley, 1996; Endsley & Kiris, 1995). In the nuclear power domain, poor situation awareness has been identified as a factor in previous plant accidents, and is thus "one of the most critical contribution[s] to safe operation" (Ha et al., 2007, p. 2692). In a nuclear plant control room, it is important to consider not only the situation awareness of the individual, but also of the operating crew, collectively (Hallbert, 1997).

As with mental workload, there are a variety of definitions of situation awareness (Sarter & Woods, 1991; Theureau, 2000). Essentially, situation awareness is "knowing what is going on around you" (Endsley, 2000, p. 5). The concept is most relevant and meaningful in terms of an operator's goals and tasks (Endsley, 2000). As defined by U.S.

Air Force Tactical Command, situation awareness includes recognizing what relevant information is unknown (McKinnon, 1986, in Taylor, 1990).

Endsley defined the construct more formally as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1988, p. 792). This definition specifies three levels of situation awareness, with each level building on the previous one (Hogg et al., 1995). The first level requires perception of individual elements relevant to the task (Endsley, 2000). In order to attain Level 2 situation awareness, the operator must integrate newly perceived information, along with information from previous assessments of the situation, into a coherent view of the system state in working memory; a lack of expertise hinders this process (Sarter & Woods, 1991). Sarter and Woods emphasized the temporal aspects of situation awareness, as the operator must retain and build on knowledge over time, rather than starting with a blank slate at each situation assessment. This implies a role of long-term memory, as well (see also Endsley, 2000).Comprehension of the current situation can also serve as a guide for future information search (Sarter & Woods, 1991). The operator may predict the future system state (i.e., achieve Level 3 awareness) from the current situation by mental simulation (Sarter & Woods, 1991); successful prediction depends on the operator having an accurate mental model of the system (Endsley, 2000; Endsley et al., 2003). Prediction of problems enables preventative actions (Sarter & Woods, 1991). The emphasis on the operator's ability to project future events in terms of task goals distinguishes situation awareness from other cognitive constructs (Durso et al., 1999). In

an experiment simulating air-traffic control, Durso et al. (1999) observed trade-offs between controller focus on present versus future awareness, suggesting that both should be considered as part of system evaluation.

Endsley's domain-independent definition of situation awareness can be applied operationally by the identification of specific task-relevant information (Endsley, 1988). Situation awareness is deemed to be multi-dimensional because the various aspects of a complex situation must be understood (Endsley, 2000; see Pew, 1995, in Hallbert, 1997 for a taxonomy of types of awareness).

Situation awareness, decision-making and performance are all related probabilistically, but these should be viewed as three separate stages. That is, a person with better situation awareness is more likely to make good decisions, and in turn, to perform satisfactorily. Other factors such as workload, stress and interface design may also contribute to the likelihood of success at each stage. In cases of remote operation, the user interface may be the operator's only source of system information, as compared to piloting or driving, for example, in which other sights and sounds provide meaningful cues. (Endsley, 2000)

Situation awareness matters during system design and evaluation because the operator, as a human information processor, is commonly viewed as being limited, perceptually and cognitively. Therefore, simply providing the operator with more information does not necessarily mean better performance (Endsley, 2000). First, the operator must selectively devote his or her limited attention to relevant information sources for a given condition (Ha et al., 2007; Wickens, 2002b). Data overload describes

conditions in which relevant information is available but is lost in the noise (Endsley et al., 2003). Attentional tunneling is another potential barrier, particularly under stressful situations, in which the operator focuses on only a portion of the available information to the point of neglecting other critical sources (Endsley et al., 2003; Hancock & Szalma, 2003). Attending to relevant information may be challenging if, as in many interfaces, some data is not always displayed, but rather available on demand (O'Hara & Hall, 1992, in Ha et al., 2007). Assuming key system information is successfully communicated, the operator must either correctly pattern-match the information to a known prototypical system schema, or interpret the information as an unfamiliar situation, and then respond accordingly (Endsley, 2000).

The recurring situation assessment process should be a combination of top-down (i.e., goal-driven) and bottom-up (i.e., data-driven) processing, with goal re-prioritization driven by emerging information (Endsley, 2000). Therefore the perceptual salience of interface components (e.g., alarms versus status indicators) is an important design consideration (Endsley et al., 2003). A high false alarm rate can result in operators disregarding even salient display elements (Endsley, 1996). Learned scan patterns may aid overall awareness by ensuring periodic monitoring of available information sources. However, the task of maintaining situation awareness is not easily proceduralized (Wickens, 2002b). Highly familiar conditions can be a double-edged sword (see, e.g., Taylor, 1990). The term automaticity describes such conditions, in which the operator may perform adequately with minimal effort, but may fail to detect irregular events (Endsley, 2000).

While there are benefits of automation in system design, this approach presents challenges to interface designers and operators in terms of maintaining situation awareness. Repetitive mental tasks such as calculations are prime candidates for automation (Endsley & Kiris, 1995). Computerized systems can process raw data and display the resulting information in higher forms more readily usable in the operator's tasks, such as trend plots (Endsley, 1996; see Spence, 2007 for an introduction to information visualization techniques).

Automation of decision-making in system control ranges from fully manual to fully automated control (Endsley, 1996). In probabilistic risk analysis, automation reduces uncertainty by eliminating the potential for unpredictable operator actions (Ha et al., 2007), and in practice, introduction of highly reliable system automation has reduced human error (Wiener, 1985, in Endsley & Kiris, 1995). However, it appears humans are not well suited for passively monitoring fully automated systems, meaning this human-machine configuration may yield suboptimal system performance (Endsley, 1996). Out-of-the-loop syndrome describes the deleterious effects of highly automated decision-making: the potential for loss of operator skills and situation awareness due to complacency or the vigilance decrement (Endsley et al., 2003; Endsley & Kiris, 1995; Hallbert, 1997). Having empirically investigated the effects of various levels of automation on operator situation awareness, Lin et al. (2010) argued against full automation. If full automation is implemented, the user interface should provide information about system actions and the motivation behind them (Mouloua et al., 2001). Complex automation poses challenges to the operator for projecting system behavior in

the near future (i.e., achieving Level 3 situation awareness; Endsley, 1996). Also, upon automation failure, it may be difficult for the operator to detect and understand the situation, and to act correctly (Endsley & Kiris, 1995; Hallbert, 1997). To keep the operator of a highly automated system "in the loop," it seems critical to design the interface and task for active, rather than passive, monitoring and to present appropriate feedback and status information regarding system behavior (Endsley & Kiris, 1995; Sarter & Woods, 1991). Wickens (2002b) called this design problem "a tremendous challenge" (p. 131; for aviation psychology, in particular). In summary, situation awareness is an especially important interface design consideration for systems featuring some level of autonomy.

2.1.5 Redlines

Given that workload assessment is an important facet of system evaluation, it is natural to ask "How much workload is too much?" (Rueb et al., 1994, p. 47). Conversely, the notion of "adequate situation awareness" (e.g., Plott et al., 2004, p. 5-13) implies a lower acceptable bound. Such a threshold for acceptable workload levels, commonly called a "redline," is particularly important in certifying a new system design (Wickens, in Grier et al., 2008). As an example, NUREG-0711 includes the notion of "acceptable workload levels" in nuclear plant designs (O'Hara et al., 2004, p. 8), while NUREG/CR-6838 warns against "excessive cognitive workload" (Plott et al., 2004, p. 5-8). Whereas relative evaluation is useful for comparing design alternatives, redlines address the problem of absolute evaluation, or determining the adequacy of a particular design (Reid

& Colle, 1988). While this concept is discussed in the literature, there are few examples of actual numbers (Reid & Colle, 1988).

Hart and Wickens (2008) defined the workload redline as the transition point from "safe and effective multi-tasking to dangerous and ineffective multi-tasking" (in Grier et al., 2008, p. 1204). In practice, this threshold should present itself as a "knee" or discontinuity in the performance curve, explainable by task shedding (Wickens, in Grier et al.). Similarly, others have placed the redline at the boundary of regions A and B in O'Donnell and Eggemeier's (1986) model of workload and performance (Colle & Reid, 2005; Rueb et al., 1994). That is, the redline identifies conditions for which workload is sufficiently high that performance may suffer. Long-term or frequent operation past such a redline increases the probability of failure and may therefore be detrimental to system performance (Colle & Reid, 2005; Reid & Colle, 1988).

System or task performance-based redlines, as specified in regulations or system requirements, clearly support design evaluation (Rueb et al., 1994). However, these should be supplemented with workload redlines, because a veteran operator may be able to support acceptable system functioning at an unacceptable workload cost (Rueb et al., 1994). One challenge is specifying a redline which is general (Wickens, in Grier et al., 2008), transferable and global (Colle & Reid, 2005), meaning it can be applied to a variety of environments and tasks. Kaber and Mosaly (in Grier et al., 2008) proposed an approach to redlines incorporating workload, situation awareness, and other cognitive factors. They suggested that such guidelines should be informed in part by data on

operational incidents (i.e., outside the lab and simulator), following the pattern of industrial ergonomics' approach to physical workload.

Since mental workload cannot be reliably inferred from primary task performance, workload redlines may be specified in terms of subjective, physiological or secondary task measures (Rueb et al., 1994). The concept of acceptable mental workload is built into the subjective Modified Cooper-Harper scale, and there is some empirical support for the implied redline based on MCH results (Colle & Reid, 2005). Based on a review of several laboratory studies applying another subjective measure, SWAT, Reid and Colle (1988) proposed a "very tentative guideline" (p. 1417) for overload prediction. Based on conjoint measurement, SWAT produces workload values ranging from 0 (minimum) to 100 (maximum). Using 10-15% error as the performance threshold, Reid and Colle repeatedly, and rather informally, found that the "critical SWAT value" for acceptable performance fell within the range of 40 ± 10. Colle and Reid (2005) later produced similar results in a simulated combat flight task. As task difficulty increased, operationalized as greater airspeed and decreasing time between targets, SWAT ratings increased linearly. However, observed performance exhibited an inverted U-shaped curve. Colle and Reid therefore termed the peak of the curve the point of "performance saturation." This point was found to be at a SWAT value of 41.1. The authors presented this as further evidence of the 40 ± 10 guideline, but noted that this was specific to the aviation domain; general applicability requires further work.

In evaluating the acceptability of a proposed crew reduction via cockpit re-design, Rueb et al. (1994) applied this SWAT redline in a simulator. They emphasized that the

SWAT redline of 40 serves as a "rule of thumb" or a "caution signal" (p. 50): cases where this number is exceeded merit further investigation. However, they found this number to be reasonable, based on their particular results. Their results also suggested potential problems with redlines due to individual differences: flight crews from different aircraft produced contrasting results. The same study proposed and employed an alternate redline technique, using the SWORD measure, in which experts were asked to judge their workload relative to that of an existing system. This technique requires veteran users of an existing system, and assumes that performance and workload are currently acceptable. Rueb et al. (1994) argued that secondary tasks are probably not suitable for redline evaluations. However, they believed that a comparison technique between the existing and re-designed interfaces using physiological measures is a promising method of evaluation.

In the only publication found to propose a numeric guideline, Reid and Colle (1988) were concerned only with overload. In the studies they reviewed, participants rarely made errors in the less demanding conditions (i.e., with low presentation rates). However, the inverted U theory of performance as a function of workload predicts diminished performance in cases of very low workload, or "underload." Braby et al. (1993) investigated underload assessment, comparing subjective and physiological (i.e., heart rate) measures in a vigilance task. Based on their results, they promoted physiological measures for detecting low arousal. Furthermore, they tested a "Subjective Work Underload Checklist," associating underload with boredom. However, not having measured performance, this study stopped short of proposing an underload redline.

We found relatively little discussion of redlines for situation awareness. As of 1995, no guidelines existed for minimal levels of situation awareness (Pew, 1991, in Endsley, 1995). However, these will be necessary if situation awareness is to be a factor in system design certification (Endsley, 1995). Just as good situation awareness does not guarantee adequate performance, good performance may possibly be achieved despite poor situation awareness (Endsley, 2000), suggesting acceptable limits for situation awareness are related to the tolerance for human error in a system (Endsley, 1995). Patrick et al. (2006) emphasized that situation awareness is "situation-specific" (p. 415). Situation awareness may thus be more meaningful in relative, rather than absolute, terms (Endsley, 2000).

2.2 Theoretical Frameworks for the Evaluation of User Interfaces

As its name implies, the field of human-computer interaction deals with the interplay of physiological, psychological, and technological concerns. User interface designs must be feasible, given certain technical constraints, but they should also support the user, catering to his goals and strengths and accounting for his weaknesses. An understanding of human perception, cognition and attention provides a solid conceptual foundation for the interface designer and a useful perspective for the evaluator of a candidate design. A comprehensive review of relevant work in psychology is out of the scope of this work. However, several noteworthy theories from psychology, human factors and human-computer interaction are briefly outlined below. These may serve as useful tools at various stages of the user interface design and evaluation process.

2.2.1 Signal Detection Theory

Signal Detection Theory (e.g., Green & Swets, 1966), as applied to psychophysics, deals with the human observer's ability to detect weak signals amidst noise. This theory provides structure for quantitatively analyzing the operator's "decisions made under conditions of risk and uncertainty" (p. 7), which is applicable in measuring the performance of monitoring tasks. In the simple yes-no task, the experimenter manipulates the presence or absence of a target stimulus in an information display, and for each presentation, the observer responds whether the stimulus was present. Within a trial there are thus four possible combinations of presence/absence and observer response; these are hit, false-alarm, miss, and correct rejection.

The sensitivity and the response criterion (i.e., the bias or willingness to respond "yes" vs. "no") of the observer can be determined from the results. Within this theory, the motivation and strategy of the operator matter, as some errors may be more costly than others. Based on waveform analysis of the background noise, and, as applicable, the signal, an ideal observer can be described which specifies the benchmark for the optimum level of performance obtainable.

Vigilance tasks, characterized by "extended periods of continuous operation" (p. 332) with a low event rate, are known to result in a performance decrement over time. Signal Detection Theory has proven useful in studies concerning this phenomenon, in showing that it is due not just to a decrease in sensitivity, but also to the development of a stricter response criterion (see Green & Swets, 1966). Accordingly, Pattyn et al. (2008)

noted that the vigilance decrement does not seem to be due solely to lapses of attention, but also potentially to changing strategies. Related to vigilance in the operational context, the Situation Awareness Control Room Inventory (SACRI; Collier & Folleso, 1995; Hogg et al., 1995) applies Signal Detection Theory to the analysis of operator responses to situation awareness queries in a nuclear power plant control room simulator.

2.2.2 Multiple Resource Theory

Multiple Resource Theory (e.g., Wickens, 2002a, 2008) explains the manner in which and predicts the extent to which performance will decrease under multi-tasking. As such, the theory is related to, yet distinct from, mental workload and the workload redline (Wickens, 2008). In contrast to previous formulations of mental capacity as a pool of general resources, this theory differentiates between various resources on several dimensions: stages of processing (perception, cognition, response), codes of processing (spatial or verbal) and modalities (auditory or visual). This theory incorporates aspects of Baddeley and Hitch's (1974) model of working memory, in which the visuospatial sketchpad and the phonological loop are fairly independent components (in Proctor & Van Zandt, 2008). A later form of Wickens' 3-D cube model of multiple resources incorporates a fourth dimension of visual channels (focal or ambient).

In this theory, dual-task performance will be superior for tasks utilizing separate resources, as they can be processed in parallel. Computational models have been developed based on Multiple Resource Theory that predict the performance decrement under dual task conditions, depending on the total resource demands and the degree to

which the tasks compete for the same resources. Such predictive models have to some extent been empirically validated. Multiple Resource Theory is also consistent with the current understanding of brain anatomy (Wickens, 2008). The theory has proven useful in human factors design for addressing overload issues under multi-tasking. While Multiple Resource Theory predicts overload conditions, it does not explain performance decrements due to underload (Wickens, 2002a).

2.2.3 Attention Investment Theory

Attention Investment Theory (Blackwell & Green, 1999) characterizes human interaction with an abstract information artifact, such as a user interface, in terms of the strategy by which the user directs attention over time. This theory is motivated in part by the understanding that attention is a valuable resource. Here, the user frequently reassesses the current course of action based on trade-offs between the estimated costs and risks of disrupting attentional focus from the current sub-task, and the anticipated benefits of doing so. Blackwell and Green noted that each context switch (i.e., change of focus) during an information transcription task incurs an attentional cost, but that the longer such necessary switches are postponed, the greater the perceived risk (cf. attentional tunneling; Hancock & Szalma, 2003). Similarly, there is a cost involved in sequentially directing attention to multiple information sources (e.g., scanning displays). This is particularly relevant to monitoring tasks with multiple modules, for which attention is at a premium, and there are trade-offs between deep understanding of a single subsystem and timely assessment of the overall system.

This work also describes affordances in terms of cognitive dimensions and attention investment. As examples, juxtaposition of related information decreases attentional costs, just as a close mapping of information formats reduces the costs and risks of transcription.

2.2.4 Information Foraging Theory and Information Scent

Like Attention Investment Theory, Information Foraging Theory (Pirolli & Card, 1995) analyzes human actions in terms of costs and benefits. Here, the cost is effort expended in interacting with (e.g., searching or browsing) an information source and the benefit is the information gathered relevant to task goals. Information Foraging Theory draws inspiration from the fields of anthropology and biology, in which Optimal Foraging Theory provides a framework for analyzing the structure of food sources and the associated food seeking behaviors. Pirolli and Card extended this theory to information sources and seekers, with the seeker (i.e., a user) making decisions on how to proceed with a system based on the value of the information available and the affordances of the system for navigating to a new "information patch." From the perspective of Information Foraging Theory, user interface design should optimize the structuring of information in support of efficient navigation and sense-making. One important observation from Pirolli and Card (1995) is that often the information seeker must first determine what questions to ask; these subsequently guide the foraging task. Also, information applicable to task goals will often be grouped together spatially and temporally, with the time for moving between such groups presenting a cost.

An extension of this work considers the relative "information scent" of user interface elements, as estimable by their perceived costs and relevance to task goals (e.g., Chi et al., 2001). Information scent shapes information foraging actions. In particular, scent applies to the information in close proximity to a navigable link, as this suggests the nature of the linked page. User actions can be modeled formally based on both information foraging and information scent.

2.3 Heuristic Evaluation of User Interfaces

Heuristic evaluation is an informal, inexpensive method for identifying usability issues in a user interface (Nielsen & Molich, 1990). This technique can be applied early in the design process with minimal effort and does not require user participation. Nielsen and Molich noted that lengthy sets of guidelines can be intimidating, and instead presented a list of nine brief rules, or heuristics, identifying common usability issues. With minimal training (e.g., a lecture covering the heuristics), a non-expert can inspect a system for such issues. In several validation studies, Nielsen and Molich found that while any one person performing heuristic evaluation could only find a portion of the known issues, good results could be obtained by aggregating the findings of three to five independent evaluators, as each person may hone in on different issues. The authors noted that their validation efforts were limited to relatively small user interfaces, and recommended that heuristic evaluation be accompanied by other evaluation, such as user studies.

Along the vein of heuristic evaluation, Annex D of IEEE Std. 1023-2004 (2005) contains a Human Factors Engineering (HFE) design review checklist intended "to direct

the attention of designers and reviewers who are not HFE specialists to common sources of HFE discrepancies" (p. 28). On a much larger scale, NUREG-0700 (O'Hara et al., 2002) contains hundreds of pages of design review guidelines for human-system interfaces in nuclear power plants. These are intended for guiding Nuclear Regulatory Commission staff in inspecting a design, but they can also inform designers earlier in the design process. The guidelines cover a broad range of topics, from general user interface components (e.g., windows, icons and cursors) to current directions in the industry (e.g., advanced alarm systems and computer-based procedures).

2.4 Experimental Evaluation of Performance, Mental Workload and Situation Awareness in User Interfaces

2.4.1 Criteria for the Selection of Experimental Techniques

A great many techniques have been developed for the experimental evaluation of user interfaces. Even if the evaluation approach is focused to assessing mental workload or situation awareness, there is still a large number of candidate techniques.

O'Donnell and Eggemeier (1986) discussed five criteria for the selection of a mental workload measure: sensitivity, diagnosticity, primary task intrusion, implementation requirements and operator acceptance. While terminology varies somewhat across sources, these are common criteria (e.g., IEEE Std. 845-1999, 1999; Zhang & Luximon, 2005). In providing guidance for human-system performance evaluation in nuclear power generating stations, IEEE Std. 845-1999 (1999) adds several additional criteria: acceptability (i.e., consensus among experts in the field), accuracy (minimal potential for error), applicability, bias, precision and reliability. Zhang and

Luximon (2005) also suggested the criterion of selectivity, favoring measures that reflect mental workload without interference from other factors such as emotional stress and physical workload. Similar criteria were found for situation awareness assessment techniques (Endsley, 1995; Salmon et al., 2009). These criteria are summarized in Table 2.1.

Diagnosticity considers the degree to which a measure can indicate the nature of the workload and the specific resources taxed (O'Donnell & Eggemeier, 1986). Validity encompasses a number of things: convergent validity (Rubio et al., 2004); face validity, construct validity, concurrent validity, and predictive validity (Salmon et al., 2009). Also, Endsley (1995) proposed the criterion that a technique not impact the construct under measurement, which is a slightly different value than low intrusiveness (of primary task performance).

Table 2.1 Review of Criteria for the Selection of Experimental Techniques.

HUMAN-SYSTEM PERFORMANCE	MENTAL WORKLOAD		SITUATION AWARENESS	
IEEE Std. 845-1999	O'Donnell & Eggemeier, 1986	Zhang & Luximon, 2005	Endsley, 1995	Salmon et al., 2009
Acceptability (consensus in field)				
Accuracy				
Applicability				
Bias				
Intrusiveness	Primary task intrusion	Intrusiveness	"does not substantially alter the construct" (p. 66)	
Precision				
Reliability		Repeatability	Reliability	Reliability
Resources (required for implementation)	Implementation requirements	Convenience (of implemenation)		
Sensitivity	Sensitivity	Sensitivity	Sensitivity	
Validity		Validity	Validity	Validity
	Diagnosticity	Diagnosticity	Diagnosticity	
	Operator acceptance			
		Selectivity	"is not a reflection of other processes" (p. 66)	

2.4.2 Performance Assessment Techniques in the Nuclear Power Domain

Specific performance measures must be selected based on the nature of the system goals, the tasks under consideration and the experimental questions. This means that

performance comparison between different candidate interface designs for a system is possible, but comparison between dissimilar systems may not be meaningful. In practice, performance measurement may require system instrumentation, such as user action logging (e.g., Ha et al., 2007), or expert observation and judgments (e.g., Ha et al., 2007; Norros & Nuutinen, 2005). Performance may be influenced by various factors, including the environment (e.g., O'Hara et al., 2004), and individual differences in training, experience (e.g., Norros & Nuutinen, 2005) and strategy (e.g., Ha et al., 2007).

In human factors for nuclear power plant control rooms, NUREG-0711 promotes performance-based evaluation of the human-system interface, as part of the design validation process (O'Hara et al., 2004). In this particular domain, human performance should be considered not just for the individual operator, but also for a control crew working together (Ha et al., 2007). With the ultimate goal being safe plant operation, simulator-based evaluation should consider whether plant parameters are maintained within pre-defined bounds (Ha et al., 2007). However, this measure may not be sensitive to the demands placed on the crew or the adequacy of the interface, as a veteran crew may exhibit satisfactory performance despite design issues (Ha et al., 2007; Norros & Nuutinen, 2005). Furthermore, the additional layers of protection afforded by safety systems (i.e., "defense-in-depth"), including passive safety features, in a modern nuclear plant imply that even human performance failures should have minimal impact on plant safety (Norros & Nuutinen, 2005). Therefore, in addition to their ultimate effects on plant safety, the individual operator's actions themselves should be evaluated: an evaluation

should measure the operator's ability, under various conditions, to detect significant events during monitoring and to respond appropriately (Ha et al., 2007).

Ha et al. (2007) proposed an evaluation of operator performance in a nuclear plant control room simulator based on a hierarchical task analysis. With this method, subject matter experts observe the operator's execution of a goal-based operating procedure, and judge the speed, accuracy, and order in which the essential subtasks are completed. This study also promoted pre-specified time limits for critical operator actions as performance requirements.

Norros and Nuutinen (2005) also proposed "a new kind of performance measure" (p. 355) which considers not just the result, but also the manner in which the task is performed. They operationalized a crew's "habits of action" in the simulator as 51 specific indicators of practice, judged by subject matter experts. Adequate task performance was defined as comparable to or better than that observed with the existing control room interface. While the habits of action analysis provided useful information, outcome-based measures of performance did not vary significantly between conditions. Similarly, Hwang et al. (2008) found task completion insensitive to various conditions, as all participants successfully completed the assigned simulated reactor shutdown task.

2.4.3 Mental Workload Assessment Techniques

The various experimental measures of mental workload can be divided into four categories: subjective, physiological, primary task and secondary task measures (Luximon & Goonetilleke, 2001; Proctor & Van Zandt, 2008; Sirevaag et al., 1993;

Zhang & Luximon, 2005). Some authors use a three-category classification, in which primary and secondary task measures are collectively termed performance-based or behavioral measures (see Baldwin, 2003; O'Donnell & Eggemeier, 1986; Veltman & Gaillard, 1996).

Zhang and Luximon (2005) rated subjective measures as the "best mental workload measures available at the present" (p. 201). These measures, which are subjective in the sense that they rely on an operator's self-assessment of workload, are relatively easy and inexpensive to implement (Sirevaag et al., 1993; Zhang & Luximon, 2005). Compared to the other categories, the need for specialized equipment and training is minimal (Baldwin, 2003), and subjective measures can be transferred to new situations and environments (Casali & Wierwille, 1983 in Svensson et al., 1997; Hill et al., 1992). Operator acceptance is typically high (Proctor & Van Zandt, 2008; Sirevaag et al., 1993), and as long as data collection can be delayed until after the experimental task is completed, intrusiveness is not an issue (Sirevaag et al., 1993; Veltman & Gaillard, 1996). However, subjective measures provide time resolution on the order of minutes at best (Svensson et al., 1997), and subjects must be instructed in the assessment technique (Baldwin, 2003).

Asking the operator to rate workload based on his or her subjective experience gives these measures relatively high face validity (Sirevaag et al., 1993; Zhang & Luximon, 2005). It is important to determine whether the operator experiences overload, even if, objectively speaking, this is not the case (Zijlstra, 1993). On the other hand, there are cognitive processes and factors influencing workload of which the subject has no

conscious awareness (Boff & Lincoln, 1988, in Proctor & Van Zandt, 2008; Svensson et al., 1997). Subjective measures of workload have been shown to dissociate from performance measures in various cases (Vidulich & Wickens, 1986; Yeh & Wickens, 1988), but this does not necessarily imply that objective measures of workload are better (Tsang & Velazquez, 1996). Also, it is possible that subjects may rate their perception of task demands rather than the resulting workload imposed (Boff & Lincoln, 1988, in Proctor & Van Zandt, 2008; Veltman & Gaillard, 1996). This class of measures is generally sensitive to changes in the overall level of workload (Sirevaag et al., 1993; Veltman & Gaillard, 1996; Zhang & Luximon, 2005), but more so in the intermediate than high workload ranges (Sirevaag et al., 1993). Uni-dimensional subjective workload measures lack diagnosticity, but multi-dimensional ones provide some level of information regarding the nature of task demands (Hill et al., 1992; Rubio et al., 2004; Tsang & Velazquez, 1996).

Colle and Reid (1998) demonstrated that subjective workload ratings are sensitive to context effects, posing a threat to the external validity of the results (i.e., the conceptual range of workload levels from "very low" to "very high" is itself subjective). This issue may be mitigated by exposing the participant to a wide range of task difficulty prior to data collection, explicitly describing extreme workload conditions, or instructing the participant to reference all previous operating experience for the range of possible workload values (Colle & Reid, 1998; Reid & Colle, 1988). Subjective measures are also not suitable for assessing operator state during operations, particularly as compared to the continuous physiological measures (Baldwin, 2003). Vidulich and Wickens (1986)

concluded that subjective measures can be "valuable," but that there may be risk in relying solely on such a measure.

Physiological or psychophysiological measures provide continuous, objective workload indications (Sirevaag et al., 1993; Veltman & Gaillard, 1996), which can be useful in assessing the operator state in real-time (Baldwin, 2003). This is potentially useful during system evaluation for linking workload variations to particular events (see, e.g., Berka et al., 2005; Tremoulet et al., 2009), and during operations for dynamically tuning the system automation level or user interface to the current demands on the operator (Berka et al., 2004). Physiological measures thus provide insight into mental demands both over the course of a task and from moment to moment (Baldwin, 2003; Proctor & Van Zandt, 2008). Lin et al. (2005) stated that subjective and performance-based measures are essential to usability evaluation, but that these two alone are insufficient, and they should be augmented by physiological assessment. Wilson (2002) concluded that "a much better picture" (p. 16) of pilot mental workload is obtained when physiological data are available.

A potential drawback of physiological measures is that they generally require costly equipment (Proctor & Van Zandt, 2008; Sirevaag et al., 1993; Zhang & Luximon, 2005), and in some cases expert analysis (Sirevaag et al., 1993; e.g. neuroergonomics per Baldwin, 2003). Some measures require complex data processing, for example to remove artifacts introduced by operator motion (e.g., EEG; Gevins & Smith, 2003). Zhang and Luximon (2005) found physiological measures "very useful" yet inferior to other methods, as they are "'polluted' by noise" (p. 201). The same authors also considered this

category of measures to be neither diagnostic nor selective. Indeed, physiological measure data can vary with physical activity (Veltman & Gaillard, 1996) and environmental factors (Zhang & Luximon, 2005).

There is disagreement on whether physiological techniques in general are intrusive (see, e.g., Brookings et al., 1996; Rowe et al., 1998; Veltman & Gaillard, 1996; but also Zhang & Luximon, 2005); this question is perhaps easier to answer based on the specific task and measure combination. Equipment which does not interfere with the operator's primary task is obviously preferable (Proctor & Van Zandt, 2008); if this is possible, then data can be obtained through physiological measures without any explicit action by the subject (Sirevaag et al., 1993).

Covariance between various physiological measures is generally low in comparison to the agreement found between several subjective measures (Svensson et al., 1997); therefore convergent validity favors the subjective measures (see, e.g., Rubio et al., 2004). O'Donnell and Eggemeier (1986) found that the various physiological measures may each be best suited to investigating particular processes; any single measure may not provide a reliable indication of overall workload or arousal.

Finally, mental workload may be inferred from operator performance, either of the primary task(s) or of a secondary task introduced for this purpose. Primary task measures give a direct indication of performance by the operator and the system overall (Sirevaag et al., 1993). The redline between acceptable workload and overload can be meaningfully defined in terms of task performance measures (Proctor & Van Zandt, 2008). The drawback of this approach is that it assumes a direct relationship between

performance and workload, whereas in reality there may be a range of mental workload levels below the redline for which primary task measures are not sensitive (Proctor & Van Zandt, 2008). That is, if the operator is not overloaded, primary task measures do not reflect the spare capacity aspect of mental workload (Sirevaag et al., 1993). An operator may tune his or her strategy to maintain acceptable performance as task demands, and presumably workload, increase (Veltman & Gaillard, 1996). Performance and mental workload are therefore different constructs, leading Zhang and Luximon (2005) to judge performance-based measures as having low validity.

Primary task measures are unobtrusive but they are not diagnostic (Proctor & Van Zandt, 2008). Depending on the measure, instrumentation may be costly (Proctor & Van Zandt, 2008). Furthermore, primary task measures are not easily adaptable to new situations (Sirevaag et al., 1993). In a multi-task condition, it is unclear how to combine multiple primary task measures into a single performance rating, as priorities may fluctuate over time (Veltman & Gaillard, 1996). Guhe et al. (2005) concluded primary task performance alone is an inadequate measure of mental workload.

Whereas primary task measures do not reflect spare mental resources, secondary tasks are intended to do precisely this (Sirevaag et al., 1993). The operator is instructed to prioritize either the primary or secondary task, and the performance of the lower priority task indicates workload in terms of spare mental capacity (Proctor & Van Zandt, 2008). Secondary task measures are thus more sensitive to workload demands in the non-overload region (Proctor & Van Zandt, 2008). Secondary tasks may also be diagnostic, as the task can be selected to reflect the reserve capacity of a particular cognitive system

(Proctor & Van Zandt, 2008). Such measures are easily transferrable to new situations (Sirevaag et al., 1993). Secondary task measures are criticized for their intrusiveness (Sirevaag et al., 1993; Zhang & Luximon, 2005), and for adding artificiality to the experimental conditions (Proctor & Van Zandt, 2008).

2.4.3.1 *Subjective Measures of Mental Workload*

A subjective measure of workload relies on the participant to reflect on his or her experience and to rate it on one or more quantitative scales. Therefore the resulting rating reflects not just the objective task demands, but also their interaction with the individual rater (Hart & Staveland, 1988). A uni-dimensional workload scale presents one continuum ranging from very low to very high workload, whereas a multi-dimensional scale prompts the participant to rate various aspects of the experience independently. A multi-dimensional measure may provide diagnosticity, as the various subscales indicate the sources of workload in a way that a single workload continuum cannot (Rubio et al., 2004; Vidulich & Tsang, 1987). The subscales may also be combined into a single, composite score for the condition. Very simple (uni-dimensional) subjective measures may be applied during task performance with minimal disruption, such as ATWIT (see, e.g., Ahlstrom & Friedman-Berg, 2006). Others place greater demands on the participant, and thus must be administered either after the task is complete, or during a task interruption (e.g., a simulator freeze). This choice may depend on the length of the condition and the nature of the research questions. Subjective techniques also vary in producing either an absolute judgment (i.e., a numeric result, possibly compared to some

standard), or a judgment relative to a baseline condition (Reid & Colle, 1988; Rueb et al., 1994; Vidulich & Tsang, 1987).

Many subjective measures have been proposed. A review of a subset of these follows. In a brief review of subjective workload assessment, Proctor and Van Zandt (2008) presented modified Cooper-Harper, SWAT, NASA-TLX and Workload Profile as "four of the most popular" measures (p. 254-255). In reviewing "state-of-the-art methods" (p. B-1) for assessing cognitive workload, NUREG/CR-6838 lists four questionnaires which can be administered during simulator freezes: Overall Workload, SWAT, NASA-TLX and the Multiple Resources Questionnaire (Plott et al., 2004, Appendix B). Simplified SWAT is an adaptation of SWAT, proposed for reducing SWAT's implementation costs (Luximon & Goonetilleke, 2001); a discussion of this adaptation is included in the section on SWAT. Finally, in reviewing the workload literature we found RSME to be heavily cited and fairly widely implemented (see analysis in section 5.1).

The Air Traffic Workload Input Technique (ATWIT) merits mention as a domain-specific, instantaneous workload measure (Ahlstrom & Friedman-Berg, 2006). Typically at five-minute intervals, the controller is prompted to manually select one of ten buttons on a continuum, reflecting his or her current workload experience. With this technique, workload is defined in practice as the task completion rate (i.e., a one on the ten-point scale represents low workload, based on full task completion). The simplicity of ATWIT implies minimal unobtrusiveness. The ATWIT method could certainly be adapted for use in other domains, but it is excluded from further investigation for the

purposes of this study. Discussion of each of the eight subjective measures of interest follows, categorized into sections as uni- or multi-dimensional techniques.

2.4.3.1.1 Uni-dimensional Subjective Measures

Overall Workload Scale

The Overall Workload scale (OW; Vidulich & Tsang, 1987) is extremely simple to administer and analyze. The scale is presented as a horizontal line with twenty intermediate demarcations and the extremes labeled as "Low" and "High." The experimenter subsequently interprets the operator's marking on the continuum as a score between 0 and 100. Given the wide consensus that workload is multi-dimensional, OW requires the participant to integrate his or her experience into a composite score. However, in a set of single- and dual-task tracking conditions, Vidulich and Tsang (1987) found that results with the OW scale were comparable to those with NASA-TLX, which has multi-dimensionality, and thus additional complexity, built-in (see below). Notably, both were apparently outperformed by a relative judgment approach (i.e., the Analytic Hierarchy Process), which asked participants to compare the workload levels in pairs of conditions. One advantage of OW over the pair-wise judgments, though, was that workload could be assessed immediately after each condition, rather than relying on retrospective estimates.

A benefit of using the Overall Workload scale, as described in NUREG/CR-6838, is that it can be administered during task performance, whereas other more complex

measures require either freezing the simulation or delaying administration until the experimental condition is complete (Plott et al., 2004). Plott et al. deemed OW to be "nearly as effective" (p. B-3) as the other subjective measures considered. Hill et al. (1992) noted the ease of application and data analysis with OW, and showed that the instrument has relatively high operator acceptance, based on explicit operator ratings. Out of the four subjective measures investigated in that study, Overall Workload also demonstrated good sensitivity, second only to NASA-TLX.

As compared to a weighted combination of multiple subscales, Hart and Staveland (1988) found overall workload ratings to exhibit high between-subject variability. They also suggested that a single overall scale reflects different factors in different conditions, such that "global workload ratings cannot be compared between tasks" (Hart & Staveland, 1988, p. 167).

Modified Cooper-Harper

As its name implies, the Modified Cooper-Harper scale (MCH; Wierwille & Casali, 1983) evolved from the Cooper-Harper scale (Cooper & Harper, 1969), a widely used rating scale for aircraft handling. Whereas the original scale deals primarily with psychomotor task load in aviation, MCH focuses on aspects of perceptual, mediational (cognitive), and verbal communications demands, and is intended for general applicability. Wierwille and Casali (1983) presented evidence from flight simulator experiments for the sensitivity of MCH to each of these three sources of mental workload. To obtain a rating of "overall mental workload" (p. 131) on the MCH scale

between 1 (“Very easy, highly desirable”) and 10 (“Impossible”) for a condition, the operator is asked to navigate a decision tree, answering questions regarding task completion, error frequency and consequences, and difficulty. The question “is mental workload level acceptable?” may imply a workload redline between MCH levels 3 and 4 (Skinner & Simpson, 2002, in Colle & Reid, 2005). Furthermore, this implicit MCH redline seems to fit well with that proposed for SWAT, based on a linear transformation of experimental SWAT results onto the MCH scale (Warr et al., 1986, in Colle & Reid, 2005).

In a study comparing four subjective workload measures, Modified Cooper-Harper was generally outperformed by the others, particularly NASA-TLX and OW (Hill et al., 1992). Although MCH could be completed relatively quickly, it had consistently lower sensitivity, as measured by factor validity. Furthermore, operators rated MCH as the most difficult instrument to complete and as an inferior description of workload.

RSME

Considering human-system interface design as an optimization problem of maximizing task performance and minimizing human cost, Zijlstra (1993) proposed and validated the Rating Scale Mental Effort (RSME) as a means of assessing the latter, particularly for human-computer interaction. He distinguished between objective workload (i.e., relatively static task demands) and the more dynamic “mental effort,” which is related to the attentional demands of a task, as well as to psychophysiological state (e.g., fatigue, circadian rhythms). Mental effort is under cognitive control; that is,

effort investment is a decision and is therefore influenced by motivation. An operator may compensate for a suboptimal state by exerting extra effort, in seeking to maintain task performance. Since mental effort can be viewed as energy expenditure, effort provides an estimate of subjective cost in task performance.

Intended for easy, low-cost application by system designers during task execution, the RSME technique presents the user with a uni-dimensional scale ranging from 0 to 150. The original form of the scale featured nine textual anchors in Dutch (e.g., at the extremes, a rating of 2 may be translated as "not effortful" and 113 as "awfully effortful"; Veltman & Gaillard, 1996). Since there is no zero-point, Zijlstra presented the scale as having interval properties. An initial laboratory experiment showed that RSME and heart rate measures responded similarly to manipulations of task demand. Furthermore, Zijlstra found that the RSME, as applied to bus drivers on their routes, was sensitive to passenger and traffic volume, as well as shift-related effects on the pschophysiological state. Zijlstra concluded that use of a uni-dimensional subjective effort scale, along with "an extensive psychological analysis of the task" (p. 53), is sufficient, as compared to a multi-dimensional scale.

RSME seems to be in fairly wide use, but more so in Europe than the United States (e.g., Veltman & Gaillard, 1996). In a simulated flight task, pilot RSME ratings did not vary significantly between tasks, whereas multiple physiological measures were sensitive to the different conditions (Veltman & Gaillard, 1996). Lin et al. (2005) adapted RSME for measuring participant stress in human-computer interaction, and found this measure distinguished between task manipulations at a significant level.

2.4.3.1.2 Multi-dimensional Subjective Measures

NASA-TLX

According to Proctor and Van Zandt (2008), NASA-TLX may be the most widely used subjective measure of workload. The NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988) was designed to provide a diagnostic summary of workload, sensitive to a variety of tasks and difficulty levels, and suitable for application during operation, while minimizing inter-individual variability. The technique was based on a conceptual cognitive model of how a human operator constructs an estimate of the workload experienced in accomplishing a task. With this index, the participant provides ratings for a task on six bipolar dimensions of workload: mental demand, physical demand, temporal demand, (estimation of own) performance, effort, and frustration level. Dimensions of stress, fatigue and activity type were considered but excluded based on initial results. Of particular note, own performance was deemed important because it impacts future effort and may reflect conditions meriting further scrutiny due to a performance decrement. Also, frustration captures the psychological effects of the task demands, and is more precise than the vaguely defined concept of "stress."

The six subscale ratings for NASA-TLX, each between 0 and 100, are linearly combined into a weighted global score. The weightings are obtained by fifteen pairwise comparisons of the dimensions: the participant must judge which of a pair contributes more to overall workload *for the task type under consideration*. Both the relative

weightings and the individual subscale ratings provide diagnostic information about the sources of subjective workload, particularly as compared to a uni-dimensional workload continuum.

NASA-TLX was the result of three years of investigation, validation and refinement, based on the results of many experiments with a variety of task categories, including supervisory control and simulated flight (Hart & Staveland, 1988). Although there was some correlation between the subscales, each was especially sensitive to certain conditions and thus provided unique information. Compared to a uni-dimensional workload measure applied in parallel, NASA-TLX was found to be more sensitive to experimental manipulations and to exhibit lower between-subject variability (but see also Vidulich & Tsang, 1987). Furthermore, the multi-dimensional measure demonstrated a test-retest reliability of 0.83. Hart and Staveland (1988) recommended instructions, examples and reference tasks for calibrating the workload ratings of individual participants into better alignment (i.e., to reduce variability).

Boles and Adair (2001) described the administration of NASA-TLX as "cumbersome" (p. 1790) and criticized its dimensions as being too general for diagnosticity. Tsang and Velazquez (1996) found that the NASA-TLX dimensions lacked a theoretical basis. However, Finomore et al. (2008) have called NASA-TLX "the current standard" (p. 1212) in the subjective measurement of mental workload for vigilance tasks, in particular. A comparison study found NASA-TLX to correlate with performance measures more highly than SWAT or WP (Rubio et al., 2004). These authors recommended applying NASA-TLX over the other two measures, in cases where the

ultimate goal is predicting task performance. Hill et al. (1992) found NASA-TLX to be more sensitive to workload variations in several army tasks than OW, SWAT or MCH. Although the NASA-TLX ratings took longer to complete than the other measures (i.e., an average of 51.3 seconds), the instrument was rated as the most likable and the best descriptor of workload.

SWAT and Simplified SWAT

The Subjective Workload Assessment Technique (SWAT; Reid & Nygren, 1988) was designed for minimally intrusive workload assessment in applied (i.e., operational) settings, and for requiring relatively simple operator judgments. Based on a literature review, the developers incorporated three commonly recognized components of workload as dimensions of SWAT: Time Load, Mental Effort Load and Psychological Stress Load. The third dimension is described as "anything that contributes to an operator's confusion, frustration, and/or anxiety" (emphasis preserved from source; Reid & Nygren, 1988, p. 191), including training, fatigue, emotional state and environmental factors.

At data collection (i.e., "event scoring"), the operator must rate each dimension from 1 (low) to 3 (high load), as guided by textual descriptions of each dimension-level pair. The three dimensions are combined additively into an overall score from 0 to 100 by a conjoint measurement / conjoint scaling approach, which requires the operator to rank the 27 possible workload rating combinations from least to greatest overall load by sorting a set of cards. One advantage of this combination method is that it produces a

scale with interval properties (Hill et al., 1992). Luximon and Goonetilleke (2001) rated the technique relatively high in diagnosticity and content validity.

Although SWAT has been criticized for practical reasons, it has been widely applied. The measure lacks sensitivity in the low workload region (Luximon & Goonetilleke, 2001). Based on their work, Hart and Staveland (1988) concluded that the three dimensions of SWAT provide inadequate coverage of the range of workload factors present in diverse task conditions. Between-subject variability is frequently high, perhaps explained by the fact that the card sorting phase is guided by the individual's general workload definition, rather than the individual's view of workload in the present context (Hart & Staveland, 1988). Also, Boles and Adair (2001) and Luximon and Goonetilleke (2001) criticized SWAT for the time required for the card sorting phase and the complexity of the conjoint scaling technique. In fact, the card sorting phase can be "quite tiring for the subjects" (Rubio et al., 2004, p. 69). Furthermore, with participants drawn from the diverse population of the U.S. army, Hill et al. (1992) observed that 43% of the time, the first card sort attempt failed to produce a valid ordering. Luximon and Goonetilleke (2001) proposed five simplified variants of SWAT that are intended to address these practical issues; all five demonstrated superior sensitivity. The adaptations appear to merge some desirable features of SWAT with the continuous subscale presentation of NASA-TLX. Nevertheless, SWAT in its original form has been used extensively (see, e.g., Svensson et al., 1997).

Of the meager research to date related to workload redlines, much of it has been carried out using SWAT (see e.g., Colle & Reid, 2005; Reid & Colle, 1988; Rueb et al.,

1994). Tentative guidelines for an upper bound on acceptable workload have been proposed in terms of SWAT, whereas no such recommendations were found for other workload measures.

Workload Profile

The eight dimensions of the Workload Profile (WP; Tsang and Velazquez, 1996) were derived directly from Wickens' Multiple Resource Theory (see, e.g., Wickens, 2008): perceptual/central and response stages of processing; spatial and verbal codes of processing; visual and auditory input; manual and speech output. Following the completion of all experimental conditions, a matrix pairing task conditions and dimensions is presented to the participant. For each task-dimension pair, the participant must rate the proportion of attentional resources used, ranging from 0 ("no demand") to 1 ("maximum attention"). Explanations and examples of each dimension are provided for reference during the rating process. In the initial validation study investigating eight task conditions, this evaluation process lasted 15 to 30 minutes for each participant. In practice, not all dimensions may apply to a particular task; two were omitted in the initial study. Although an overall workload score was obtained as an unweighted sum of the dimensional ratings, the authors admitted this approach may not be ideal and expressed uncertainty regarding the best way of computing a composite score.

Tsang and Velazquez (1996) emphasized the diagnosticity of their Workload Profile instrument. In their study using tracking and memory tasks in single- and dual-task conditions, not only did the dimensional rating results (i.e., the workload profile)

reflect pre-experiment expectations of the task demands, they also deepened the understanding of the task demands by their quantitative nature. The authors therefore concluded that participants were able to meaningfully self-assess demands on the various dimensions included. Although high inter-dimensional correlation was also observed, Tsang and Velazquez proposed that diagnosticity at the level provided by WP may be sufficient to pinpoint opportunities for improvement in training, procedures or the user interface. The authors noted that for diagnosticity, the WP may be an attractive alternative to intrusive secondary tasks.

In the initial validation, WP demonstrated high test-retest reliability, but its sensitivity to task manipulations was slightly inferior to two uni-dimensional measures: Psychophysical Scaling and the Bedford Scaling procedure. In contrast, a comparison study by Rubio et al. (2004) found WP to outperform NASA-TLX and SWAT in terms of sensitivity and diagnosticity. In this study, WP ratings were collected after each individual condition. Rubio et al. (2004) also reported instances of participants struggling to understand the WP dimensions; notably, their subjects were university psychology students.

Multiple Resources Questionnaire

The Multiple Resources Questionnaire (MRQ; Boles & Adair, 2001) is a relatively recently proposed multi-dimensional measure of workload, designed especially for identifying mental resource conflicts during multitasking. The questionnaire seeks to measure the extent to which seventeen mental processes are taxed in a task, with each

item ranging from 0 (no usage) to 4 (extreme usage). The seventeen dimensions were identified based on Multiple Resource Theory, as expanded by Boles' previous work (Boles, 1991, 1992, 1996, and Boles & Law, 1998, in Boles & Adair, 2001). The authors noted that not all dimensions may apply for a certain task, and may possibly be excluded from the questionnaire. For example, "short-term memory process," "spatial attentive process" and "manual process" are clearly relevant in operator interaction with a real-time critical user interface, but recognition of faces or facial emotion may not be. An alternative is to present all seventeen but to omit from analysis dimensions that received a score of zero by more than half of the raters (Finomore et al., 2008). The instructions emphasize that the participant should rate the average level of process usage over the course of the task, not the peak demand. Unlike other multi-dimensional workload measures, there is no proposed procedure for calculating a single composite workload score.

The authors promoted the diagnosticity of MRQ in terms of specific mental processes, as compared to the more general, high-level dimensions of other measures such as NASA-TLX and SWAT. In computer game task-based evaluations, they typically found inter-rater reliabilities around $r = 0.6$. They also found that some respondents apparently failed to read and follow the directions. Of the subjective workload measures, MRQ is relatively wordy.

MRQ has proven useful in demonstrating that the workload of vigilance or sustained attention is relatively high (Finomore et al., 2008). In this study, each of the seventeen dimensions ranged from 0 to 100, and the mean of the relevant dimensions for

a given condition was reported as a global workload score. Not only did MRQ produce results comparable to NASA-TLX (including some significant inter-instrument correlation, particularly with the mental demand subscale of NASA-TLX), it also identified multiple loaded mental processes that NASA-TLX does not consider. However, the authors found it inconvenient to compare global workload ratings from MRQ between dissimilar conditions, because the number and type of relevant contributing subscales (i.e., the loaded resources) varied. They also noted that it is geared toward visual processes and thus lacks diagnosticity for auditory tasks.

2.4.3.2 *Physiological Measures of Mental Workload*

This section describes various physiological measures used to assess workload. Rowe et al. (1998) categorized such physiological measures into two classes: "central nervous system," including brain and electro-oculographic activity, and "peripheral nervous system" measures.

Here, the measures are grouped into the following sub-categories by physiological system: neuroergonomic, cardiopulmonary, eye-based measures and galvanic skin response. This classification is useful for organizational purposes. More importantly, if multiple physiological measures are to be selected (see e.g. Hwang et al., 2008 for motivation), it may be desirable to select representative measures from multiple sub-categories.

2.4.3.2.1 Neuroergonomic Measures

Neuroergonomics is an interdisciplinary field applying neuroscience, the study of the nervous system, to the human factors domain (Parasuraman & Wilson, 2008). This includes a set of psychophysiological techniques for investigating and assessing mental workload (Parasuraman & Wilson, 2008). Baldwin (2003) stated that this approach to mental workload measurement is "in its infancy" (p. 134), but it potentially provides for "a more direct assessment of mental workload" (p. 134), as compared to other methods. These techniques also offer superior diagnosticity, as workload variation may be localizable to particular structures within the brain (Baldwin, 2003; Fournier et al., 1999; Wilson, 2002; see O'Donnell & Eggemeier, 1986, for a general discussion).

While equipment to date has been unwieldy and data analysis complex (Baldwin, 2003), both event-related analyses and frequency domain analyses of electroencephalographic (EEG) data have proven useful in assessing workload. According to Parasuraman and Wilson (2008), EEG-based approaches provide good temporal resolution and relative ease of use for human factors purposes, but they lack spatial resolution in terms of the brain's topography. Gevins and Smith (2003) also found the cost of EEG measurement equipment to be low. According to Berka et al. (2005), "EEG is the only physiological signal that has been shown to reflect subtle shifts in alertness, attention and workload that can be identified and quantified on a second-by-second basis" (p. 91). Other neuroergonomic methods, such as positron emission tomography and magnetoencephalography (PET, MEG; see Rowe et al., 1998),

functional Magnetic Resonance Imaging (fMRI; see Parasuraman & Wilson, 2008) and functional Near-Infrared (fNIR) have also been proposed, but are at present not practical for realistic usability testing and will not be discussed in detail.

EEG data is collected by placing multiple electrodes on the participant's scalp, where the brain's electrical impulses can be measured (O'Donnell & Eggemeier, 1986). The International 10-20 System specifies a standard set of sensor locations from which the experimenter may choose (see, e.g., Sirevaag et al., 1993). Traditionally, a wire connects each sensor independently to the data acquisition equipment. Setup can be tedious, and data collection may be hindered by intrusion of the equipment, depending on the nature of the task. These problems may be alleviated through the use of fitted caps with embedded sensors placed based on approximate head size (see, e.g., Berka et al., 2004). Furthermore, a wireless headset featuring signal digitization near the sensor addresses the traditional problem of cable noise from electromagnetic interference, and may eliminate the need for scalp preparation (Berka et al., 2004).

EEG-based mental workload assessment faces several other practical barriers besides bulky equipment. Large individual differences are found in the data (Baldwin, 2003; Sirevaag et al., 1993). Also, application outside of controlled lab environments is currently difficult because EEG signals are influenced by factors besides mental workload such as eye blinks, eye movement and motor activity (Baldwin, 2003; Gevins & Smith, 2003). Periods of EEG data with such artifacts must be identified for exclusion or correction during data analysis; the latter is preferable in the case of frequent perceptuomotor activity (Gevins & Smith, 2003). Clearly, technical expertise is required

for proper experimental setup, data collection, processing and analysis. In fact, Farmer and Brownson (2003, in Hwang et al., 2008) deemed EEG impractical as a mental workload measure for these and other reasons.

Various Event-Related Potentials (ERP)s have been identified in EEG signals, each featuring a characteristic signature and latency in response to some experimental event. In a review of mental workload measures, Proctor and Van Zandt (2008) rated ERPs as the most reliable neuroergonomic measure of brain activity, as associated with certain processes. One ERP commonly used for workload assessment is the P300 (or P3), so named because it occurs approximately 300ms after stimulus onset (O'Donnell & Eggemeier, 1986). This time period of the "transient evoked response" is commonly associated with information processing and cognitive activity in general (O'Donnell & Eggemeier, 1986).

The participant is instructed to watch or listen for a pre-defined, low-probability stimulus, in addition to performing the primary task. When the stimulus is detected, the P300 should be subsequently evoked. The response for each of many trials must be checked for undesirable artifacts. Digital filtering and range correction may be employed (e.g., Sirevaag et al., 1993). Because the signal-to-noise ratio for ERPs is relatively low, many trials are typically averaged for data analysis (Gevins & Smith, 2003; O'Donnell & Eggemeier, 1986; Proctor & Van Zandt, 2008; for an example, see Sirevaag et al., 1993). The P300 method is particularly suited to assessing the workload demands of monitoring for "rare or novel stimulus events" (Proctor & Van Zandt, 2008, p. 254). Researchers may also make judgments on workload demands in general based on the latency and

amplitude of the P300 (O'Donnell & Eggemeier, 1986; Proctor & Van Zandt, 2008). However, previous work indicates that P300 may be more sensitive to perceptual task demands than response selection difficulty (O'Donnell & Eggemeier, 1986).

P300 latency is known to increase as the difficulty of detecting the pre-defined target increases (Parasuraman & Wilson, 2008). Latency of this ERP has also been shown to increase as the primary task becomes more complex (e.g., Kramer et al., 1987 in Proctor & Van Zandt, 2008). Also, in a simulated helicopter flight task, Sirevaag et al. (1993) demonstrated that amplitude of the P300 response to an irrelevant auditory probe decreased with greater communication task loads. However, P300 amplitude also decreases with greater expectancy (O'Donnell & Eggemeier, 1986), and therefore after repeated presentations (Proctor & Van Zandt, 2008). O'Donnell and Eggemeier (1986) suggested using primary task-related stimuli, such as system auditory alarms, for evoking P300 unobtrusively during operations.

A second common EEG-based approach to measuring mental workload analyzes signals in the frequency domain (i.e., spectral power analysis). Recent work in this area demonstrates the potential applicability of such an approach to continuous, real-time assessment of mental workload during operations (Berka et al., 2004, 2005; Gevins & Smith, 2003; Parasuraman & Wilson, 2008). One advantage of this approach is that no task-irrelevant stimulus is required. Multiple studies have found that power in the alpha band (8-12 Hz) of human brain activity tends to decrease with greater workload (Brookings et al., 1996; Fournier et al., 1999; Gevins & Smith, 2003; Wilson, 2002). Meanwhile, power in the theta band (approx. 4-7 Hz) tends to increase with task

difficulty (Brookings et al., 1996; Gevins & Smith, 2003; but see Wilson, 2002). In reviewing work on theta power, Brookings et al. (1996) noted that increased theta activity may indicate fatigue, vigilance decrement or higher cognitive activity.

Having first shown the theta and alpha bands sensitive to working memory load manipulations, Gevins and Smith (2003) proceeded to demonstrate a continuous (0.25 Hz) index of mental workload during human-computer interaction based on artificial neural network pattern classification of multiple EEG variables. Their approach mitigated the effects of individual differences, as the classifier was trained to the individual operator prior to data collection. They also addressed the issues of artifacts by introducing algorithms for auto-detecting and "decontaminating" such undesirable periods from the data via adaptive filtering (see also Fournier et al., 1999 regarding correction of eye movement periods). Their results provide evidence for selectivity of the spectral power measure, as the workload estimates produced were more sensitive to variations in cognitive demands than in perceptuomotor task demands.

Berka and colleagues (2004, 2005) developed the B-Alert system, a real-time classifier of "alertness, cognition and memory" (2004, p. 151), which uses data from three channels of EEG acquired with a wireless headset. The raw data is automatically pre-processed to eliminate artifacts resulting from eye blinks and operator motion. The system classifies the moment-to-moment state of the user at 1 Hz as one of four (2004) or five (2005) levels based on an analysis of nineteen variables, including factors of alpha, beta and theta power. Notably, this approach assumes that mental workload measurement is synonymous with assessing the operator's functional state on a continuum of

drowsiness to high vigilance. The classification model is tuned to the individual based on three baseline conditions. The model has been validated with various types of tasks (Berka et al., 2004). A follow-up study simulating an Aegis (naval ship) monitoring and response task demonstrated high temporal correlation between the B-Alert workload measure and event-driven, demanding operator tasks (n = 5; Berka et al., 2005). One favorable result of this approach is the ability to characterize a task by the percentage of total time an operator spends at each level of workload.

Tremoulet et al. (2009) extended this work, presenting a tool called SMART for evaluating user interface designs in terms of workload. SMART includes the B-alert wireless EEG headset and classification software. This study found statistically significant correlation between their "domain-independent, task-independent cognitive workload measure" (p. 342) and HCI expert ratings of workload. Interestingly, correlation of SMART's index with NASA-TLX subjective ratings only approached significance. Also, the overall mean estimate of workload produced by SMART varied little between experimental phases.

Showing promise in lab and simulated settings, neuroergonomic mental workload assessment is an active research area with potential applicability to adaptive automation, adaptive user interfaces and dynamic task loading within a crew during operations (e.g., Baldwin, 2003; Berka et al., 2005; Parasuraman & Wilson, 2008; Tran et al., 2007b). However, further technical advances are needed to address the practical issues of noisy data and intrusive equipment before this goal can be achieved (Parasuraman & Wilson, 2008).

2.4.3.2.2 Cardiopulmonary Measures

Cardiopulmonary measures provide an indication of workload. Several of these have been applied to the assessment of mental workload in particular, including heart rate (HR), heart rate variability (HRV), blood pressure (BP: systolic and diastolic), blood pressure variability (BPV), and various measures of respiration. Blood volume measures have also been investigated (Rowe et al., 1998). Cardiac measures have traditionally been the "most popular physiological techniques" (p. 481) for assessing mental workload (Rowe et al., 1998). In the literature, respiration and cardiac measures have proven sensitive to mental workload in certain cases, but they generally exhibit low selectivity. That is, factors besides mental workload can certainly affect these cardiopulmonary measures.

Heart Rate and Heart Rate Variability

One common way to measure heart rate is via electrocardiogram (EKG or ECG), in which electrodes placed near the heart or, alternatively, at the sternum and side of the ribcage, register the heartbeat signature known as "QRS" (Brookings et al., 1996; Fournier et al., 1999; O'Donnell & Eggemeier, 1986; Sirevaag et al., 1993). Inter-beat interval can be automatically determined by the time between consecutive R-peaks, but this technique may require human verification (see, e.g., Sirevaag et al., 1993; Veltman & Gaillard, 1996). EKG is generally seen as unobtrusive (O'Donnell & Eggemeier, 1986).

Multiple studies have found heart rate to be sensitive to workload in simulated environments. Fournier et al. (1999) found EKG, including heart rate, along with performance-based measures, to be the most effective of many workload measures in a multi-tasking experiment. In a flight combat simulator with active military pilots, significant positive correlation was found between mean heart rate per mission and mission difficulty; mean heart rate and information complexity on a heads-down tactical display; mean heart rate and SWAT (Svensson et al., 1997). This same study found that the running mean of the heart rate was especially sensitive to information load with expert pilots. Veltman and Gaillard (1996) also found mean inter-beat interval (i.e., essentially the inverse of heart rate) per segment of a simulated flight task to reflect varying task demands. Hwang et al. (2008) found heart rate to increase with task difficulty for most novice operators in a simulated reactor shutdown task. They therefore included heart rate in their workload/performance prediction model incorporating multiple physiological measures.

Other studies have not found heart rate useful as a mental workload indicator. In investigating novice operator mental workload with human-system interfaces for advanced nuclear plants, Jou et al. (2009) used mean heart rate per condition minus resting heart rate as a physiological measure. They did not find a significant difference in heart rate between any of the conditions in a 2 x 2 experimental design varying plant automation level and reactor operator task. In a simulated air traffic control task with experienced controllers, heart rate was not sensitive to variations of traffic volume or complexity (Brookings et al., 1996). Similarly, Sirevaag et al. (1993) did not find

significant differences in mean inter-beat interval between conditions in a helicopter simulator experiment.

Furthermore, compared to physical exertion and arousal level, mental workload is not the primary determinant of heart rate (Proctor & Van Zandt, 2008); that is, heart rate has low selectivity. O'Donnell and Eggemeier (1986) stated that heart rate is "probably not useful as a workload measure" (p. 42-39) due to its sensitivity to other factors, but they did place more confidence in heart rate variability (HRV), at least for tasks with low motor demands. Sirevaag et al. (1993) took the following somewhat conflicting stance: while heart rate seems to reflect overall mental workload, HRV, when analyzed in the frequency domain, may be sensitive to particular types of demands. Compared to some other measures, Rowe et al. (1998) found EKG-based HRV measurement favorable in terms of sensitivity, diagnosticity, unobtrusiveness and convenience. In a review of the literature, they found HRV is of value in both lab and operational contexts.

According to Veltman and Gaillard (1996), however, heart rate variability has proven more effective for workload assessment in the laboratory than in the field. In their flight simulator-based study, HRV as a workload measure was confounded by respiration. The authors therefore recommended that respiration should be incorporated in studies using HRV as a mental workload measure. Sirevaag et al. (1993) shared this concern, cautioning that differences in the amount of speech (and thus respiration behavior) between conditions may interfere with the HRV measure.

Nonetheless, some studies have found that HRV tends to decrease with greater task demands (Fournier et al., 1999; Hwang et al., 2008). Tharion et al. (2009) observed

lower HRV in university students on the day of an examination compared to HRV during vacation and attributed the difference to "anxiety-induced mental stress" (p. 63). With novice participants in an air traffic control-like display monitoring task, HRV was a marginally significant indicator of task demands (Rowe et al., 1998). However, with experienced controllers performing the same task, HRV decreased significantly with greater task demands up to a point, and then subsequently increased with greater task load. The authors speculated that this "knee" in the HRV curve (see Wickens in Grier et al., 2008) may reflect the point at which task demands became too great and resulted in task shedding. That is, they have potentially identified a physiological measure-based "redline" for this particular task. Rowe et al. therefore concluded that HRV merits further study as a measure of mental effort, both for the evaluation of user interfaces, and as an input to adaptive interfaces.

As with EEG data, power spectral density analysis of heat rate may provide additional insight into mental workload (Rowe et al., 1998). By applying a Fast-Fourier Transform to the heart rate data, Svensson et al. (1997) found power in the low and intermediate frequency bands to diminish with higher information load. In simulated air traffic control, HRV as a workload indicator approached significance with varying task difficulty, but only in the 0.15-0.4 Hz band (Brookings et al., 1996). HRV is widely reported as sensitive to workload around 0.10 Hz (i.e., in the low frequency or LF range 0.04-0.15 Hz; O'Donnell & Eggemeier, 1986; Proctor & Van Zandt, 2008; Rowe et al., 1998; Sirevaag et al., 1993; Svensson et al., 1997; Tharion et al., 2009). Healey and

Picard (2005) used the LF/HF (low frequency to high frequency power) ratio as an indicator of stress in a real-world driving task.

Kalsbeek (1973, in O'Donnell & Eggemeier, 1986) found that over thirty unique techniques have been used to determine heart rate variability. This great diversity of calculation methods is a possible explanation for the mixed results obtained with HRV in the literature. According to Healey and Picard (2005), while a 300 second window for HRV calculations is typical, their results suggested a 100 second window may be adequate.

Blood Pressure and Blood Pressure Variability

Elevated blood pressure is an indication of sympathetic (i.e., "fight-or-flight") response to the present conditions (Veltman & Gaillard, 1996). While there are multiple reports of blood pressure's sensitivity to workload level, there is relatively little discussion of blood pressure variability. Veltman and Gaillard (1996) used a tonometer with a finger cuff to measure participants' blood pressure in a simulated flight task. Although this method was subject to artifacts due to unintentional finger motion, both systolic and diastolic blood pressure varied predictably, and significantly, between segments of the flight: blood pressure was highest during flight with a secondary continuous memory task, and during the landing phase. Blood pressure variability was also measured, but it was deemed insensitive to task demands.

Via digital sphygmomanometry, Tharion et al. (2009) found that mean arterial blood pressure was significantly higher in students on the day of a university exam, as

compared to during vacation. Finally, in a simulated reactor shutdown task, Hwang et al. (2008) used an ambulatory blood pressure monitor designed for use with an upper arm cuff. They found that systolic blood pressure was a significant predictor of performance in novice participants, but diastolic pressure was not. However, they also reported that for many individuals, blood pressure did not increase significantly with greater task complexity.

We suspect that these measures may be comparable to the other circulatory measures in terms of face validity and selectivity. While blood pressure has demonstrated some sensitivity, we encountered relatively few studies which have applied this measure. Therefore it may merit further study in the lab prior to adoption for applied efforts.

Respiration

In a widely cited article, Wientjes (1992) discussed the methods of respiration measurement and their application to psychophysiology. Wientjes found that modern methods are practical, unobtrusive and reliable, but also noted that respiration is highly variable by behavior and by the individual. Involving time and volume components, respiration as a physiological measure is multidimensional and thus complex. In summarizing previous work, Wientjes found that respiration has proven useful in many studies on mental demand and stress. In the literature, "rapid shallow breathing with a high inspiratory flow rate" (p. 196) is often linked to mental effort, sustained attention, anxiety, fear and tension (Wientjes, 1992). In terms of the measured data, mental effort or stress are typically found to be reflected by higher respiration rates and V_{MIN} (i.e., minute

volume, the total air volume in one minute of breaths), as well as decreased respiration variability and V_T (i.e., tidal volume, the volume of a single breath). (Wientjes, 1992)

For the sake of assessing workload, one common method of measuring respiration consists of placing elastic transducer bands about the subject's chest and abdomen (e.g., Brookings et al., 1996; Fournier et al., 1999; Wientjes, 1992). This approach requires calibration for the individual's volume to motion ratio, and Wientjes (1992) noted that measurement error grows over time due to slippage of the bands. Therefore frequent recalibration is suggested (Wientjes, 1992). Also, the data is rendered unusable during periods of operator speech and movement (Wientjes, 1992; see also Sirevaag et al., 1993), so this approach may have limited applicability depending on the prescribed operator tasks.

Applied studies using the elastic bands technique in simulated Air Traffic Control (Brookings et al., 1996) and with the Multi-Attribute Task Battery (Fournier et al., 1999) have found respiration rate, but not breathing amplitude, to reflect increased workload demands. Based on a review of previous work, Brookings et al. (1999) suggested that respiration may be a more sensitive measure than heart rate. However, Wientjes described a previous study (Wientjes et al., 1986, in Wientjes, 1992) in which respiration rate was not sensitive to task demands, but tidal volume increased (i.e., became less shallow) with increasing task demands. Also, in a simulated flight task with varying difficulty, Veltman and Gaillard (1996) found that the only significant difference between respiration spectral energies was in comparing the post-landing with other conditions, as breathing became slower and deeper upon conclusion of the tasks. Due to noisy data they

were unable to calculate respiratory volume. Tharion et al. (2009) found no significant difference in respiration rate in university students between vacation and examination conditions, although other physiological measures were sensitive to the varying demands.

Notably, Wientjes (1992) also suggested hyperventilation as an indication of unacceptably demanding or stressful tasks and "potentially dangerous aspects of the task environment" (i.e., a workload redline; p. 194). Hyperventilation is a "passive coping response" (p. 192) to situations in which the subject feels threatened or unable to meet the demands present. Hyperventilation can be inferred from estimates of arterial pCO_2 levels, using either infrared gas analysis of exhaled air, or transcutaneous measurement by a heated sensor placed on the skin (Wientjes, 1992).

2.4.3.2.3 Eye-Based Measures

Eye-based measures, particularly eye-tracking methods, are applicable to the evaluation of user interfaces in more ways than one: they can indicate workload levels, but they may also provide an understanding of the user's need for and access of information (Tran et al., 2007a; see also O'Donnell & Eggemeier, 1986). Various properties of human eye activity have been used to measure workload, including pupil dilation (i.e., the "task-evoked pupillary response"; TEPR) and properties of blinks, fixations and saccades. Ahlstrom and Friedman-Berg (2006, Appendix A) provided the technical definitions they used for data analysis of these properties.

Various techniques have been used to measure eye activity. In a combat flight simulator-based study, Svensson et al. (1997) were particularly interested in the duration and frequency of fixations between the heads-up cockpit view and a heads-down display. They were able to measure these relatively simple variables by manually analyzing video of pilot task performance. Many studies have recorded eye blinks and vertical motion via electrooculographic (EOG) sensors placed above and below one eye (e.g., Brookings et al., 1996; Fournier et al., 1999; Sirevaag et al., 1993; Veltman & Gaillard, 1996). An electrode placed to the side of the eye can support horizontal motion analyses (e.g., Brookings et al., 1996; Fournier et al., 1999). This technique is fairly unobtrusive and convenient (O'Donnell & Eggemeier, 1986). However, calibration needs and relatively low accuracy are potential drawbacks (O'Donnell & Eggemeier, 1986).

Head-mounted eye trackers capture the blinks and movement of one eye, and also the pupil diameter (see, e.g., Ahlstrom & Friedman-Berg, 2006; Iqbal et al., 2004; Tungare & Perez-Quinones, 2009). In a preliminary description of ongoing work, Tungare and Perez-Quinones (2009) reported using this approach to track participant eye activity across multiple devices. They found it "permits free head movement" and provides "reasonable accuracy" (p. 3434; Tungare & Perez-Quinones, 2009). Klingner et al. (2008) disagreed, judging head-mounted eye trackers as precise but prone to slippage, "cumbersome and annoying" (p. 64). Although Iqbal et al. (2005) relied on head-mounted eye tracking in assessing mental workload during naturalistic HCI tasks, they anticipated that future advances in eyeglasses- or LCD monitor-based (i.e., "remote") eye

tracking solutions would enable the practical application of eye tracking to supporting users in performing computer tasks.

Remote eye trackers require no physical contact with the user (Tran et al., 2007a), and thus potentially offer high operator acceptance and unobtrusiveness. They are also able to track both eyes. Klingner et al. (2008) replicated the results of previous pupil dilation experiments with a remote eye tracker. In doing so, they demonstrated the suitability of remote video eye tracking systems for assessing workload via the TEPR. However, their method relied on auditory stimuli with the participant's eyes fixated at the center of the monitor. Furthermore, they explicitly mentioned that eye tracker-based pupil measures are noisy, particularly those from a remote system. They therefore applied a low-pass filter to the data prior to analysis. In a more naturalistic task, Hwang et al. (2008) used eye blink frequency and duration as measures of mental workload in a simulated reactor shutdown task.

These various technical solutions contrast in intrusiveness, in cost, and in the ability to accurately measure the various properties of eye activity. A discussion of the various specific measures of eye activity, and the applicability of each to assessing mental workload, follows.

Blink Rate and Duration

Ahlstrom and Friedman-Berg (2006) defined a blink as "complete or partial closure of the eye" (p. 634). Workload studies have considered blink rate (or frequency), duration and amplitude. In an early review of workload assessment techniques,

O'Donnell and Eggemeier (1986) stated that blink rate measures "do not appear promising" due to "great variability" (p. 42-39). Fournier et al. (1999) found blink rate sensitive to task demands with the Multi-Attribute Task Battery (MATB), but it also varied with training whereas other workload measures did not, suggesting blink rate may have reflected a change in participant strategy. In a simulated air traffic control task, Ahlstrom and Friedman-Berg (2006) found no significant correlation between the number of aircraft in the sector and blink rate, although other measures were sensitive to aircraft quantity. Veltman and Gaillard (1996) did find that the interblink interval (i.e., the inverse of blink frequency) increased significantly during landing in a simulated flight task. However, there was seemingly no difference in interblink interval between flight conditions with or without a secondary, auditory task. They cited numerous studies suggesting eye blinks reflect "visual load" rather than mental workload. Brookings et al. (1996) found blink rate decreased with greater workload, but they also attributed this to increased visual demands, in particular. Sirevaag et al. (1993) found a marginally significant difference ($p < 0.09$) in blink rate between helicopter simulator conditions reliant on digital (i.e., visual) and spoken communications.

Based on their own findings and a review of the literature, Sirevaag et al. (1993) concluded that while blink rate seems to reflect visual workload, blink duration may be an indicator of both visual and auditory task demands. Blink duration was more sensitive than interblink interval to task demands in a simulated flight task (Veltman & Gaillard, 1996). Similarly, Ahlstrom and Friedman-Berg (2006) found that blink duration decreased with more aircraft present, whereas blink rate was unaffected. In the same

study, blink duration was significantly longer when air traffic controllers were provided with a weather forecasting tool in stormy conditions, suggesting workload was alleviated from the control condition (i.e., no forecasting tool). Notably, the subjective measure ATWIT did not detect a significant difference in workload between the conditions; blink duration had superior sensitivity in this case. Fournier et al. (1999) found significant differences between single- and multi-task conditions with the MATB for blink duration, as well as blink amplitude. O'Donnell and Eggemeier (1986) reported that blink duration variation in a longer term evaluation may reflect changes in workload, motivation or fatigue. The faceLAB 5 eye tracking system considers blink information in assessing fatigue and vigilance levels (Seeing Machines, 2009), just as Berka et al. (2004, 2005) equated workload with a vigilance-drowsiness continuum in classifying EEG data.

Pupil Diameter (TEPR)

Termed the "Task-Evoked Pupillary Response" (TEPR), pupil diameter has been shown to respond to task manipulations associated with mental workload (Klingner et al., 2008; O'Donnell & Eggemeier, 1986). Pupil dilation may vary closely in time with task difficulty by up to 0.5 or 0.6 mm (Klingner et al., 2008; O'Donnell & Eggemeier, 1986). Iqbal et al. (2004) found pupil size to be "the most promising single measure of mental workload" (p. 1477), as it reflects "processing load" in real-time and is minimally intrusive. TEPR may be useful in predicting performance decrements due to "sustained attention or workload" (O'Donnell & Eggemeier, 1986, p. 42-38); it may also be reliable enough to compare relative workload across tasks (O'Donnell & Eggemeier, 1986).

O'Donnell and Eggemeier (1986) viewed TEPR as "one of the most valuable indices of cognitive workload when used properly in the laboratory," and thus recommended it as the "foundation of a physiological, laboratory-based mental-workload assessment screening" (p. 42-39).

While pupil diameter is highly sensitive to mental workload, it provides minimal diagnosticity, and its low selectivity presents challenges for use outside of the lab (O'Donnell & Eggemeier, 1986). Workload effects on pupil dilation may be overwhelmed by ambient illumination effects, screen brightness and "emotional effects" (Hasett, 1978 in O'Donnell & Eggemeier, 1986; Iqbal et al., 2004; Klingner et al., 2008). Ahlstrom and Friedman-Berg (2006) also noted large individual differences in baseline pupil diameter, which may present challenges.

In a simulated air traffic control task, controller pupil size was significantly greater with a static weather forecasting tool, as opposed to a dynamic forecasting tool (Ahlstrom & Friedman-Berg, 2006). This suggests that the static tool imposed greater workload. Notably, the subjective measure ATWIT was not sensitive to this difference in task demands, suggesting TEPR had superior sensitivity in this case. Also, pupil dilation correlated positively with the number of aircraft present in the simulation.

Iqbal et al. (2004) found significantly different pupil size between various tasks, as well as high temporal resolution of the measure, corresponding to subtasks defined in a hierarchical GOMS model of the task. However, average percent change in pupil size (PCPS) per condition did not distinguish between difficulty manipulations of tasks, whereas performance-based and subjective measures did. Without averaging, though,

PCPS provides a subtask resolution not feasible with various other measures (Iqbal et al., 2005). The latter study determined PCPS at a high sample rate, from an at-rest baseline condition for two interactive HCI tasks. Pupil diameter was thus demonstrated sensitive via a variety of analyses. Although the authors found pupil size to be a highly useful measure of mental workload, they lamented the effort required for interpreting the data obtained via a commercial eye tracker.

Lew et al. (2008) investigated the utility of frequency domain analysis of pupillometric data, for the sake of assessing mental workload in augmented cognition systems. The results suggested that spectral power may increase on a broad range of frequencies following incident detection, reflecting increased mental workload or stress.

Fixations and Saccades

A fixation is defined as when the user's gaze is confined to a small area, uninterrupted by blinks or saccades (Ahlstrom & Friedman-Berg, 2006). A saccade is a quick change in eye direction to a new fixation point (Ahlstrom & Friedman-Berg, 2006). Fixation suggests the user is visually attending to the area. Fixation information therefore offers interpretability insofar as it provides the ability to link a measure of workload with the display elements receiving attention (Ahlstrom & Friedman-Berg, 2006).

Both fixations and saccades have been investigated to a lesser extent than eye blink techniques as workload measures. In general, higher *perceived* (i.e., subjective) workload is associated with operator reliance on fewer components of the display and longer fixations (O'Donnell & Eggemeier, 1986). These authors believed such scan

pattern information may indicate overall (i.e., perceptual and cognitive) workload, but diagnosticity is low.

In a simulated low-level, high-speed flight task, Svensson et al. (1997) investigated "pilots' strategies and mental workload" (p. 370) by measuring fixation durations and frequencies with various levels of information complexity presented on a heads-down (HD) display. Fixation durations on the HD display increased with greater display complexity, whereas the mean duration of heads-up fixations decreased to a certain extent. Meanwhile, the frequency of heads-up fixations decreased with greater HD display complexity.

In a simulated air traffic control task with adverse weather conditions, Ahlstrom and Friedman-Berg (2006) did not find correlation between saccade frequency and the number of aircraft present. However, they presented evidence supporting saccade distance as a physiological workload measure, as mean saccade distance decreased with more aircraft present. They further ensured that this effect was due to workload variation and not to display properties by demonstrating that the mean distance between aircraft did not decrease as aircraft quantity increased in the sector.

2.4.3.2.4 Galvanic Skin Response

Galvanic skin response (GSR), measured in siemens (Levin et al., 2006), is also referred to as electrodermal activation (EDA) or skin conductivity (Healey & Picard, 2005). Stimulation of the sympathetic nervous system can result in heightened sweat

gland productivity, and thus, increased electrical conductance of the skin (Healey & Picard, 2005; Levin et al., 2006). Various properties of GSR can be analyzed (e.g., Wilson, 2002). A sudden rise in skin conductance, indicating an autonomic response, is termed an "orienting response" (Healey & Picard, 2005, p. 161). GSR is "highly responsive," (Lin et al., 2005), as its latency following a specific stimulus is approximately 1.4 seconds (Lockhart, 1967, in Healey & Picard, 2005).

Galvanic skin response has been shown to vary with arousal, anxiety, stress, emotional response and cognitive activity (see Lin et al., 2005 for a brief literature review). This response is typically measured by running a small current through two electrodes placed slightly apart on the participant's hand or foot (e.g., Healey & Picard, 2005; Lin et al., 2005; Wilson, 2002). Guhe et al. (2005) integrated sensors for physiological measures, including GSR, in a trackball-based "emotional mouse," as part of a non-intrusive workload assessment methodology.

Various studies have investigated the relationship between galvanic skin response and workload or stress. In a small pilot study (n = 3), Guhe et al. (2005) reported that GSR decreased with decreasing inter-stimulus intervals in an auditory task (i.e., GSR decreased with increasing task difficulty). In investigating the workload of emergency department physicians, Levin et al. (2006) sampled GSR with a wireless armband at one minute intervals. They found no consistent relationship between GSR and either objective or smoothed subjective estimates of workload.

However, multiple studies have found galvanic skin response to correlate positively with stress or workload. Lin et al. (2005) found GSR to be sensitive to

variations in stress induced by three different video game tasks: GSR increased with task difficulty. In their study, mean normalized GSR also correlated well with subjective stress level, as measured by a modified RSME scale. Interestingly, there was also significant correlation between expertise and mean normalized GSR, as the top performers exhibited the lowest (normalized) skin conductance. In one condition, orienting responses (i.e., a more than 5 percent increase in GSR) were observed for 90 percent of task failures, or "frustration events."

In a real-world driving task, Healey and Picard (2005) tested GSR, along with other physiological measures, as indications of workload-induced stress. Of the measures investigated, they found GSR to correlate most highly from second to second with stress indicators on the road and in the car, as coded from video of the driving trials. They concluded that "skin conductivity is the best real time correlate of stress followed by the HRV and heart rate measures," and that between heart rate and skin conductance, "a reliable metric can be obtained" (p. 163).

Similarly, Wilson (2002) found electrodermal activation, averaged over two-minute segments of a specified flight plan, to be sensitive to task demands. Statistically significant peaks in the mean EDA responses were apparent for the take-off, touch-and-go, and landing segments of the flight. Further analysis eliminated motor activity as a potential cause of such EDA variation. In contrast to the findings of Lin et al. (2005), this study did not find much resemblance between GSR or heart rate and subjective ratings of mental workload. The author speculated this may be because the flight segments producing high physiological indications of workload were more familiar to the pilots

than some instrument flight rules (IFR) tasks. Having found a correlation of r=0.83 between GSR and heart rate, Wilson concluded that there is redundancy between the two measures, and thus that recording the more sensitive of the two, heart rate, may be adequate.

2.4.3.3 *Performance-Based Measures of Mental Workload*

2.4.3.3.1 Primary Task Measures

Primary task measures infer the workload level from the participant's ability to perform his or her assigned tasks (O'Donnell & Eggemeier, 1986). This approach assumes that as task difficulty is increased, the task demands on the operator increase accordingly, thus producing a performance decrement (Rubio et al., 2004). However, at low levels, workload variation may have little effect on performance because resources exceed demands, and the operator is able to maintain satisfactory performance (i.e., reserve capacity is not an issue; O'Donnell & Eggemeier, 1986). Similarly, at sufficiently high workload levels, further increases will likely not result in a significant performance decrement as demand already exceeds resources (i.e., reserve capacity is depleted). That is, primary task measures may be most sensitive to workload variation in Region B of O'Donnell and Eggemeier's (1986) workload-performance model (see section 2.1.3 of this report): as workload increases in this region, performance is expected to decrease accordingly. For conditions producing workload in regions A or C, on the other hand, primary task-based results may be misleading (O'Donnell & Eggemeier, 1986).

O'Donnell and Eggemeier (1986) noted that performance of the primary task can be a useful measure of workload: given a benchmark for performance, this measure can distinguish between the non-overload and overload regions. However, primary task measures are neither diagnostic nor are they necessarily transferrable or comparable across systems (O'Donnell & Eggemeier, 1986; Sirevaag et al., 1993).

Primary task-based workload assessment has been widely applied, but has produced mixed results in practice. In a multi-tasking study using multiple workload assessment techniques, Fournier et al. (1999) found task performance to be one of the most sensitive measures. In contrast, primary task performance was not a useful measure in a study by Hwang et al. (2008): all participants completed a simulated reactor shutdown task satisfactorily, despite workload manipulations. With multiple measures of primary task performance, there is high face validity and a greater likelihood of sensitivity, but this also implies more researcher effort, and presents issues of how to combine measures into a composite score (O'Donnell & Eggemeier, 1986). Rubio et al. (2004) deemed it "prohibitively difficult" (p. 63) to evaluate workload via primary task performance with complex, highly automated, modern systems. Similarly, Ha et al. (2007) believed that this approach to workload assessment is not compatible with the monitoring and decision-making tasks characteristic of an advanced reactor control room. Jones and Endsley (2004) concluded that for their study, "as with many domains, sensitive and meaningful performance measures are difficult to find" (p. 362).

In conclusion, performance-based user interface evaluation is extremely useful because meaningful benchmarks can be established. However, relying on primary task

performance as an indicator of workload may be challenging or even inappropriate for the domain under consideration.

2.4.3.3.2 Secondary Task Measures

As discussed above, measures of primary task performance may suffer from low sensitivity at certain levels of workload. Secondary task measures boast superior sensitivity (O'Donnell & Eggemeier, 1986), as long as they place demands on some of the same processing resources required by the primary task (Proctor & Van Zandt, 2008). Secondary task measures are based on measuring performance of the primary and secondary tasks individually and concurrently to infer reserve capacity under various conditions (O'Donnell & Eggemeier, 1986).

There are various formulations of the secondary task, which vary in terms of which task the operator is instructed to prioritize (for an overview, see O'Donnell & Eggemeier, 1986). The basic idea of the subsidiary task technique is to assign the operator an additional, lower priority task and to observe the effect of experimental manipulations on secondary task performance, thereby inferring the reserve capacity under the primary task. As workload increases, the operator should sacrifice performance of the secondary task to maintain satisfactory performance on the primary task. This can be highly diagnostic if the secondary task targets a particular cognitive subsystem (cf. Multiple Resource Theory; e.g., Wickens, 2008), and thus measures workload demands

in terms of resource conflicts. For a broader profile of the task demands, a set of diverse secondary tasks can be applied simultaneously (Proctor & Van Zandt, 2008).

Secondary task measurement has been widely used (O'Donnell & Eggemeier, 1986). The main criticism is that it can be highly intrusive, in that it may impact primary task performance (O'Donnell & Eggemeier, 1986). The artificial demands placed on the operator could also threaten system performance in the field and ecological validity in the simulator. This problem can be mitigated if the secondary task has relevance, yet relatively low priority, within the environment (i.e., the secondary task is "embedded;" Shingledecker et al., 1980, in O'Donnell & Eggemeier, 1986). O'Donnell and Eggemeier (1986) recommended secondary tasks which impose continuous, rather than intermittent, demands on the operator.

Because secondary tasks may be designed to target particular types of task demands (e.g., a visual stimulus detection task for measuring the visual perception demands in a system monitoring task), individual techniques are only briefly reviewed here. Jones and Endsley (2004) measured response time to verbal and visual cues. This technique was not sensitive to differences between wartime and peacetime scenarios, whereas NASA-TLX was. In a helicopter simulator study, Sirevaag et al. (1993) were able to infer workload differences between conditions based on the relative levels of shedding of embedded communications tasks. In a reactor shutdown task, Hwang et al. (2008) varied the complexity of both primary and secondary mental arithmetic tasks, and found that secondary task performance suffered at the higher workload levels, while the primary task was satisfactorily completed by all participants. Furthermore, error rates on

the secondary tasks correlated positively with NASA-TLX. O'Donnell and Eggemeier (1986) discussed common secondary tasks, and the practical aspects of administering them.

2.4.4 Situation Awareness Assessment Techniques

Although the term situation awareness (SA) has no unitary meaning (Patrick et al., 2006), it is measurable and can be a meaningful indicator of the adequacy of interface designs and training programs (Endsley, 2000). More than thirty assessment techniques have been proposed (Stanton et al., 2005 in Salmon et al., 2009), many of which are intended for simulated conditions (Endsley, 1996). There is no consensus on a single best approach (Salmon et al., 2009; Theureau, 2000). As with mental workload assessment, there are objective, subjective and indirect performance-based measures of situation awareness (Endsley, 1996). The techniques may be classified into four categories: direct query and questionnaire, physiological measurement, subjective rating (i.e., self-report or self-rating), and implicit performance-based (Durso et al., 1999; Ha et al., 2007, as derived from previous classifications; see also Salmon et al., 2009); the first two are objective techniques. Various techniques may be compatible or may have contrasting theoretical grounds. The goals and needs of a particular study should drive the selection from the various categories of measures (Endsley, 2000). For the domain of supervisory control in particular, Patrick et al. (2006) noted that situation awareness is meaningful in terms of task goals and the current conditions, so an assessment methodology should

address these explicitly. One or more prominent examples from each category of measures are described below, along with the strengths and weaknesses of each.

2.4.4.1 *Query and Questionnaire-based Measures of Situation Awareness*

SAGAT

The Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1988) is an objective measure which asks operators about critical, dynamic aspects of the system, and compares their responses to recorded system values. SAGAT is intended for application during randomly timed, brief freezes of simulated operation, with the system displays temporarily blanked. It is thus generally not suitable for use in the field. For a specific instance, either all of or a subset of the parameters deemed relevant to situation awareness are randomly selected; this provides a snapshot of the operator's current situation model. The question set addresses all three levels of Endsley's definition of situation awareness (i.e., perception, comprehension, projection; Endsley, 1996). A global situation awareness score can be calculated for each administration.

SAGAT was designed specifically for evaluations in a combat flight simulator, but has been widely used and has been applied to other domains, such as air traffic control (Endsley et al., 2000; Plott et al., 2004). While the response accuracy to particular questions provides good diagnosticity (see, e.g., Endsley et al., 1998), average SAGAT scores usually do not distinguish between conditions and may not even be meaningful (i.e., the sensitivity of a composite SAGAT score is very low; Jones and Endsley, 2000; see also Endsley et al., 2000). On the other hand, Salmon et al. (2009) found overall

SAGAT scores to correlate significantly with performance in a military planning task, for which good performance should indeed have been linked to the level of situation awareness.

The specificity of the SAGAT queries may provide information on trade-offs between various interface designs, in terms of their particular affordances and hindrances (see, e.g., Endsley et al., 1998). Jones and Endsley (2004) included a map for use in some SAGAT queries, which appears to address the spatial aspects of situation awareness more directly (see also Figure 1 in Endsley, 1995). Such visuospatial queries could potentially be applied to other domains.

While noting SAGAT's high face validity, Endsley (1988) conceded that the primary drawback of this method is the need for simulator freezes. SAGAT is criticized as being highly intrusive, potentially affecting performance or situation awareness (e.g., Ha et al., 2007; Sarter & Woods, 1991). Sarter and Woods (1991) pointed out that SAGAT measures what pilots can recall when prompted out of context, rather than what they are aware of in the moment. While freezing the simulation intrudes into primary task performance, it may not impact the construct under measurement, which is a somewhat different criterion proposed by Endsley (1995). In a pilot study, SAGAT freezes did not affect subjective situation awareness or performance (Endsley, 1988). Endsley (1995) found no noticeable decline in fighter pilot response accuracy to SAGAT queries for up to five or six minutes of freeze time, suggesting that the information was retrievable from long-term memory. A second experiment from the same study found no significant differences in pilot performance between trials with or without freezes. Patrick et al.

(2006) believed that the measure may be too global, in the sense that random freezes with non-event-specific queries may fail to capture the quality of situation awareness in the context of particular events. Salmon et al. (2009) used a Hierarchical Task Analysis for identifying situation awareness requirements for SAGAT in a military planning task. They concluded that in cases where the information critical to SA can be pre-specified and the course of events is predictable (e.g., in the simulator), "SAGAT is the most suitable approach for assessing SA" (p. 499).

SACRI

The Situation Awareness Control Room Inventory (SACRI; Collier & Folleso, 1995; Hogg et al., 1995) is an adaptation of SAGAT for nuclear power plant crews, designed particularly for evaluating interface designs in a Pressurized Water Reactor control room simulator. In devising the inventory, the developers considered situation awareness to be "the operator's overview of the current process state" (Hogg et al., 1995, p. 2395), and sought to address all three levels of Endsley's (1988) definition. Each administration consists of twelve (Hogg et al., 1995) or eighteen (Plott et al., 2004) randomly selected multiple-choice questions regarding process parameters in the form of: trends for the recent past, projections into the near future, or comparisons to typical values under normal conditions. For example, rather than reporting the current cold leg temperature value in the primary circuit, the operator is asked whether the value is greater than, less than, or the same as the value expected under normal status. The questions were developed with assistance from PWR operators and other subject matter experts. Signal

Detection Theory is applied for scoring the operator's situation awareness. Because the questions cover a wide range of parameters and the freeze points and content of a SACRI administration are randomized, the operator cannot prepare for the inventory beyond carrying out the primary task of monitoring (Hogg et al., 1995). These questions are supplemented with others capturing the operator's view of the situation with respect to task goals (Collier & Folleso, 1995).

As a result of validation efforts, the developers of SACRI found that novice operators lacked the expertise necessary for meaningful system evaluation (Hogg et al., 1995). Also, in practice, they concluded that scenarios intended for SACRI assessments should exceed 30 minutes. Collier and Folleso (1995) found the measure sensitive to individual differences in competence, to variations in an alarm interface and throughout the course of a simulated process disturbance.

SACRI inherits SAGAT's built-in freezes, and thus the associated criticism (Ha et al., 2007; Theureau, 2000). Hogg et al. (1995) found operators typically took three to seven minutes to complete the twelve questions of one SACRI questionnaire. However, operators may not object to the freezes, because they are accustomed to them as part of normal simulator-based training (Hogg et al., 1995). According to NUREG/CR-6838, "based on the experience of the Office of Nuclear Regulatory Research and others, SACRI represents an appropriate measure for assessing situation awareness" (Plott et al., 2004, p. B-2). SACRI's developers concluded that the inventory should be used as a supplement to performance measures in the evaluation of control room interfaces (Hogg et al., 1995).

SPAM / Real-time probes

The Situation-Present Assessment Method (SPAM; Durso et al., 1999) is an objective, query-based technique for measuring situation awareness which does not require freezes or rely on operator memory. With this method, also referred to as "real-time probes" by Jones and Endsley (2000, 2004), questions regarding present and future system information are presented to the operator during the task. In contrast to freeze methods, the information displays remain available to the operator during a situation query, and situation awareness is judged based on response latency rather than accuracy. In theory, a prompt response to a real-time probe indicates that the information is available in working memory (Jones & Endsley, 2004).

This technique further assumes that "SA may sometimes involve simply knowing where in the environment to find a particular piece of information" (Durso et al., 1999, p. 1-2). Queries may be built into naturalistic dialogue in order to maintain realism and minimize intrusiveness, as called for by Sarter and Woods (1991). For example, Durso et al. (1999) presented spoken queries to participants during simulated air-traffic control as telephone calls from another (simulated) location. Together with the mental demand subscale of NASA-TLX, SPAM showed some ability to predict controller performance, as judged by a subject matter expert. Also, the measure was found to be a marginally significant predictor of remaining action count at the end of a scenario: the longer it took a controller to respond to a future-related query, the less efficient his or her performance tended to be on the primary task.

One criticism of real-time probes is that they may artificially direct the operator's attention to key information, thus biasing the results (Endsley, 1995). Despite concerns that SPAM's response time-based technique may indicate workload more than situation awareness, Jones and Endsley (2000) found a weak correlation between SPAM and SAGAT, suggesting SPAM indeed measures situation awareness. With the frequency at which probe queries were presented in that study, participants did not mind the intrusion. In fact, in a related study, air traffic controllers provided a mean rating of 2.2 out of 10 for SPAM's intrusiveness (Jones & Endsley, 2004). However, the investigators recommended a higher probe frequency for the sake of improving sensitivity. While they found some cause for concern as to the validity of the measure (2004), Jones and Endsley concluded that real-time probes are "a potentially useful and effective measure of SA" (2000; p. 245) in cases when simulation freezing is not possible.

2.4.4.2 *Physiological Measures of Situation Awareness*

There is relatively little work applying psychophysiology to the measurement of situation awareness (Wilson, 2000). Wilson reviewed the use of psychophysiological techniques in other domains, speculating how these might be applied or adapted to situation awareness assessment. Such methods would be unobtrusive and would provide continuous indications. Wilson laid out several possible ways in which physiological measures may prove useful in assessing situation awareness and identified a variety of measures for consideration (e.g., heart rate, EEG, EDA, ERP/ERD, and eye activity

measures). These measures may prove effective in assessing Levels 1 and 2 of Endsley's (1988) situation awareness definition, but a psychophysiological solution for assessing situation awareness at Level 3 (i.e., the ability to predict system behavior into the near future) is not obvious. Physiological measures may prove useful in identifying the operator's functional state in relation to states conducive to achieving situation awareness. For example, fatigue and overload likely hinder good situation awareness. It should be possible to measure via physiological indicators when an operator expects an event or detects a relevant stimulus, but it is not currently feasible to measure how the operator interprets perceived information. Wilson proposed developing classifiers and training them to detect the level of situation awareness based on physiological data, and also using transitory physiological events (e.g., heart rate peaks) to identify opportune times for situation awareness probes.

Endsley (1995) discussed EEG and eye tracking as potential measures of situation awareness, but concluded that overall this category of measures is "not very promising" (p. 66). As with mental workload assessment, P300 approaches require averaging of the data from multiple stimulus presentations (Wilson, 2000). Smolensky (1993) called for investigation of eye movement measures for assessing aspects of situation awareness. Noting that most information in the nuclear plant control room is communicated visually, Ha et al. (2007) selected eye tracking as a secondary measure of situation awareness in evaluating an advanced control room design. Similar to Wilson, Ha et al. noted that eye gaze does not imply that the operator understood the information attended, it only indicates where he or she looked (i.e., it identifies "areas of interest"). The resulting data

requires expert evaluation and is time-consuming (Ha et al., 2007). Also, eye tracking techniques do not account for perception of elements in the visual periphery (Endsley, 1988). In discussing the results of their study on situation awareness in a nuclear plant control room, Hogg et al. (1995) remarked that eye tracking information could have been useful.

2.4.4.3 *Subjective Measures of Situation Awareness*

SART

Based on a multi-step process of knowledge elicitation from Royal Air Force crews, Taylor (1990) developed the Situational Awareness Rating Technique (SART). SART is a multi-dimensional, subjective assessment technique intended for aircrews, but it is also generally applicable. SART does not ask the subject to rate awareness of specific content directly. Instead, Taylor identified three fairly independent dimensions related to attention and cognition, which indicate the relative level of situation awareness: (1) demands on attentional resources, (2) supply of attentional resources, and (3) understanding of the situation. Ratings may be obtained for these three during operation. If greater diagnosticity is desired, Taylor proposed ten specific sub-areas within the three higher categories, respectively: (1) instability, complexity, variability; (2) arousal, concentration, division of attention, spare capacity; (3) information quantity, information quality, familiarity. The ten-dimension version is more suitable for post-test than real-time assessment. Taylor discussed ways of combining the various dimensions into a composite score, but did not prescribe a particular method.

Subjective techniques are attractive for their low cost, practical convenience and non-intrusive nature (Ha et al., 2007). Although SART has been used extensively (Endsley et al., 1998; Plott et al., 2004), there is disagreement regarding the validity of subjective situation awareness techniques. Generally, such subjective measures do not correspond well with performance measures, meaning that an operator may report good situation awareness, when objectively, his or her awareness must be inadequate (Plott et al., 2004). In one comparison of measures study, subjective ratings on the "understanding" dimension of SART correlated positively with subject matter expert ratings of participant situation awareness (Endsley et al., 2000). In another evaluation, SART ratings appeared to be unrelated to the objective measure SAGAT (Endsley et al., 1998). However, the authors believed SART may provide a useful indication of the subject's *confidence* in his or her situation awareness, which can impact the decision-making process, regardless of the actual, objective level of situation awareness (see also Endsley, 1995). Salmon et al. (2009) did not find SART to correlate with either performance or SAGAT in a military planning task. They concluded these two methods measure different things and they ultimately questioned the validity of SART.

By incorporating supply and demand of attentional resources, SART considers workload to be a major contributing factor to situation awareness, which may be undesirable if the two constructs are largely independent (see, e.g., Endsley, 1993; Jones & Endsley, 2000; Salmon et al., 2009). Also, in simulated air-traffic control, participants seemed to interpret the "understanding" dimension of SART as understanding of the present (i.e., Level 2 SA), which would imply that this measure does not incorporate the

future component of situation awareness (Durso et al., 1999). In relying on correct recall, post-task assessment is subject to memory decay and is therefore not appropriate for long tasks (Jones & Endsley, 2004; Plott et al., 2004). Furthermore, subjective assessment may be biased by the performance level obtained, or in cases when an operator does not recognize that key information is unknown (i.e., he or she may fail to recognize a lack of awareness; Jones & Endsley, 2004). Post-task subjective assessments may be prone to overgeneralizations (Endsley, 1995).

KSAX

Ha et al. (2007) selected KSAX (Cho et al., 2003, in Ha et al.), an adaptation of SART, for subjective situation awareness assessment in an advanced nuclear power plant control room simulator. KSAX addresses the three levels of Endsley's situation awareness definition explicitly, while omitting the workload aspects of SART (Ha et al., 2007). The KSAX questions elicit ratings relative to previous operator experience in an existing plant. This measure is noteworthy in that it has been used recently in the nuclear power industry.

2.4.4.4 *Performance-Based Measures of Situation Awareness*

Situation awareness may be inferred from operator behavior or task performance. For example, Alexander et al. (2000) recorded pilot conformance to more than 125 rules of engagement in simulated combat as a measure of situation awareness. In practice, the use of performance measures for this purpose may be difficult (Durso et al., 1999). High-

level performance measures lack sensitivity and diagnosticity, whereas a small set of low-level ones may lack the breadth for overall assessment (Endsley, 1995). Furthermore, situation awareness, decision-making and performance are all related, but these are purported to be distinct stages of information processing, meaning performance provides only an indirect measure of situation awareness (Endsley, 2000). In other words, poor performance does not necessarily indicate poor situation awareness (Endsley, 1995). Observer rating techniques can be applied in real-world settings and are unobtrusive, but it is debatable whether subjective assessment of situation awareness based on expert observation of operator actions and dialogue is valid (Salmon et al., 2009). With performance-based techniques, it may not be possible to measure situation awareness at Level 3 of Endsley's (1988) definition (i.e., projection of future state; Durso et al., 1999).

Patrick et al. (2006) implemented observational assessment of team situation awareness in three challenging nuclear plant control room scenarios. The experiment included time limits for awareness of specific conditions (e.g., each simulator crew should detect a certain incident within thirty seconds). Three experimenters with sufficient understanding of the industry rated the crews on various seven-point scales, based on crew communication, actions and the information displays available. This methodology is unobtrusive and may be superior to other approaches at addressing experimental questions regarding situation awareness under particular conditions of interest. Patrick et al. presented their methodology as practicable, and generally found good inter-rater reliability, but also noted that "considerable effort" (p. 410) was required. This observation-based measure successfully identified conditions in which situation

awareness varied between crews, as well as conditions for which all crews exhibited similar levels of awareness, whether good or poor.

2.4.5 Relating Mental Workload and Situation Awareness

Mental workload and situation awareness have frequently been discussed together, and there are multiple studies in which the two constructs were measured and compared (e.g., Alexander et al., 2000; Durso et al., 1999; Endsley, 1993; Endsley & Kiris, 1995; Hallbert, 1997; Jones & Endsley, 2004; Lin et al., 2010). Theureau (2000) noted that the two are sometimes classified similarly, because of their practical relevance and the challenges they pose to definition and measurement. In user interface and task design, there may theoretically be trade-offs between mental workload and situation awareness: mental workload may be alleviated at the expense of diminished situation awareness, just as design efforts to support good situation awareness may engender significant operator workload (Alexander et al., 2000; Endsley, 1993; Endsley, 2000). Therefore, as recommended by NUREG-0711 (O'Hara et al., 2004), both should be assessed as part of an interface evaluation effort (Alexander et al., 2000; Endsley, 2000). Good interface design may be characterized by supporting high awareness at relatively low levels of workload (e.g., the aim of an advanced alarm system in a nuclear plant control room, Hogg et al., 1995; Endsley, 1993). Automated processing and integration of system information, as well as advanced techniques of information presentation may aid operator situation awareness while simultaneously reducing workload (Hallbert, 1997). Taylor (1990) remarked that if operators will be expected to make decisions

amidst uncertainty, good situation awareness is likely a higher design priority than optimal workload.

In the literature, the composite picture from various theories and discussions regarding the relationship between the two constructs is a complex one. While maintaining situation awareness is itself a task that creates workload (Wickens, 2002b), the decisions arising from good situation awareness may initiate additional tasks, and thus additional workload (Endsley, 1988). Situation awareness impairment can be expected in the overload region (Endsley, 2000). Underload conditions may also hinder situation awareness, due to the vigilance decrement or boredom (Endsley, 2000; Hallbert, 1997). Although these are "interrelated concepts" (Wickens, 2002b, p. 128), Endsley (1993) concluded that the two may be largely independent, with the possibility for interaction in exceptional cases. She thus presented a two-dimensional theoretical continuum of situation awareness and mental workload, in which various combinations are conceivable. In an ideal system, the operator would maintain high situation awareness at a moderate level of workload (Alexander et al., 2000). Wilson (2000) linked the two constructs in a discussion of interpreting operator psychophysiological data.

The results of experiments examining the interaction between mental workload and situation awareness are mixed. In a simulated combat flight experiment, Endsley (1993) found no significant correlation between SWAT (subjective workload) and SAGAT (objective situation awareness) across all subjects. However, two subjects demonstrated negative correlation between the measures: lower workload corresponded with higher situation awareness, and vice versa. Alexander et al. (2000) measured both in

a combat flight simulator and found negative correlation between the two constructs for all seven participants; this study used the SWORD measure of mental workload and inferred situation awareness from conformance to the rules of engagement. Jones and Endsley (2004) found a significant negative correlation between subjective workload (NASA-TLX) ratings and expert observer ratings of operator situation awareness ("Observer-SART"), as well as marginally significant positive correlation between NASA-TLX and real-time probe response latency (i.e., higher subjective workload was associated with lower situation awareness, as measured by SPAM).

In simulated air-traffic control, Durso et al. (1999) found that a combination of workload and situation awareness measures predicted expert evaluation of controller performance, but that of the two, situation awareness was a superior predictor, meaning it is a construct separate from workload. In various scenarios in a nuclear plant control room simulator, Hallbert (1997) observed that following the onset of a major plant disturbance, situation awareness, as measured by SACRI, decreased, while subjective workload typically rose dramatically (it almost doubled). In these scenarios, situation awareness gradually recovered, but workload remained high as the crew worked to recover from the incident, suggesting a complex relationship between the two. Furthermore, at times of high workload, situation awareness was moderated by the level of team interaction. Seeking high ecological validity, Lin et al. (2010) recently investigated the effects of various levels of automation in a PC-based reactor simulator on operator performance, mental workload and situation awareness. From the four conditions they considered, they found that mental workload, as indicated by the mental

demand dimension of NASA-TLX and the response time in a secondary task of signal detection, decreased with greater levels of automation. Situation awareness, as measured by post-experiment queries addressing the three levels of Endsley's definition, was maximized with an intermediate level of automation termed "blended decision-making;" situation awareness levels 2 and 3 were particularly impaired at the extremes of manual and automatic control.

For studies assessing both constructs, it is desirable time- and budget-wise to select measures which are compatible within a single trial (e.g., Ha et al., 2007). As an example, NUREG/CR-6838 noted that NASA-TLX and SACRI can be administered together during a simulation freeze (Plott et al., 2004). This technique was employed in a study of control room alarm system design, as reported in NUREG/CR-6691 (O'Hara et al., 2000). Also, of particular relevance to the present research, NASA-TLX and SACRI were utilized in a study of control room staffing levels for advanced reactors (NUREG/IA-0137, Hallbert et al., 2000).

3. ANALYSIS OF CASE–SPECIFIC PROBLEM CONSTRAINTS

The selection of measures should be grounded in a solid understanding of the pros and cons of each candidate technique, as well as the requirements, constraints and needs of the particular study. The previous section described many possible measures in detail.

To tune our general recommendations to the needs of NuScale Power's Human Factors Engineering (HFE) program, we became acquainted with multiple relevant NUREGs (regulatory guidance documents from the U.S. NUclear REGulatory Commission) and recent work in the field of HFE for nuclear power plant control and monitoring. Furthermore, we tried to capture NuScale Power's initial formulation of the concept of operations, control room environment, Human-System Interface (HSI) design and operator tasks. This was accomplished by inspecting NuScale Power system design documentation and through multiple discussions with Ken Harris, the Instrumentation and Controls (I&C) Manager at NuScale Power. We also consulted with Steven Blomgren, an I&C Specialist at NuScale Power, who is tasked with the development of the NuScale simulator. NuScale is in the initial system design phase, so I&C details will certainly evolve moving forward, based in part on iterative design and evaluation. This section is intended to capture the best current understanding of the constraints for an HSI evaluation. This includes physical layout and technological aspects of the NuScale plant control room, function allocation between automated systems and the crew, and the nature and frequency of the tasks assigned to reactor operators.

This analysis and discussion begins with a brief overview of traditional control rooms and current trends, including some first-hand observations. This provides a basis for understanding future directions in human factors for "advanced control rooms," such as the NuScale Power design.

3.1 Traditional (Analog) and Hybrid (Analog and Digital) Control Rooms

The most recently built nuclear power plants currently in operation in the United States were constructed nearly thirty years ago (Werner, 2010). The main control rooms for plants from this era feature a host of analog controls and displays. Traditionally a crew is assigned to a single reactor, and by law, at most two reactors may be controlled from a single control room (Persensky et al., 2005).

As technology has rapidly advanced over the past several decades since such facilities were designed and built, some plant control rooms have undergone technological upgrades and retrofitting. These are called "hybrid" control rooms, as they feature a mix of the original analog and more modern digital equipment (O'Hara et al., 2004). Such changes potentially affect crew workload, situation awareness and performance, and must be evaluated from a human factors standpoint prior to deployment (O'Hara et al., 2004). Although we did not have the opportunity to visit a nuclear power plant control room or control room simulator during the course of this research effort, we did observe several different control rooms first-hand.

Figure 3.1 depicts a process control room at the Coors Brewery in Golden, Colorado. This configuration typifies hybrid control rooms, as it contains both legacy benchboards with analog controls and indicators, and groups of LCD displays. Some of the analog instrumentation is configured in a "mimic" fashion, meaning that the layout and visual connections communicate the relationships between system components.

Figure 3.1. Coors Brewery process control room, Golden, Colorado, visited September 2009. A hybrid (analog and digital) control room.

In February 2010 we visited the Radiation Center on the Oregon State University campus, which houses the Department of Nuclear Engineering and Radiation Health Physics, as well as several test facilities. We were granted access to the main control room of the university's TRIGA reactor while the reactor was shut-down (Figure 3.2).

This system, used for research purposes, operates at much lower power and is much simpler than a commercial power-generation nuclear plant. Accordingly, the control room and procedures are considerably simpler. However, the equipment is akin to that in a traditional plant control room (Figures 3.2-3.5).

Although the arrangement and presentation of the alarm tiles within the alarm annunciator panel appeared confusing to us (Figure 3.3), the operators stated that these indicators are grouped logically and their positions and meanings allow for quick recognition of issues. They also informed us that the digital chart plotter in the control room (center of Figure 3.4) is easier to read than analog displays. Startup and shutdown of the TRIGA are the most demanding phases of operation. Startup can take up to an hour with the TRIGA reactor, whereas this procedure takes several days in a nuclear power plant. From a design perspective, the crew emphasized the importance of spatially separating routine and critical controls, to avoid inadvertent actuation of the critical ones. It can be a challenge to keep operators engaged during normal operations.

Figure 3.2. Oregon State University TRIGA reactor control room.

Figure 3.3. Detail. Control room alarm annunciator, positioned above control area.

Figure 3.4. Detail of control area.

Figure 3.5. Detail. Analog controls and indicators.

Our visit also included a tour of the APEX test facility and control room (Figure 3.6). The facility houses a scale model of the Westinghouse AP1000 reactor and is used to test safety systems as part of the design certification process (see, e.g., NUREG-1826, Welter et al., 2005). In a typical test an incident is pre-configured and initiated, and then the researchers observe and ensure proper system response without operator intervention. The purpose of the controls and indicators on the control panel and displays is thus very different from those in an operational plant. This more modern design exemplifies hybrid qualities, and like the brewery control room, uses a mimic layout for linking related components. Our host emphasized that in monitoring such systems, there is a hierarchy of parameters critical to maintaining situation awareness. For example, core temperature,

water level, pressure and neutron flux are all top-level indicators of system performance. That is, the operator's attention should not be evenly distributed between all inputs, but should instead be prioritized based on an in-depth understanding of the system.

The philosophy of the AP1000 and other contemporary designs is for the system to respond to incidents via passive systems and highly automated actions. That is, the system should not rely on the operator to make key decisions quickly for safe operation. In fact, requirements now specify that the system should provide a 72-hour window following the onset of a disturbance, in which no operator intervention is required.

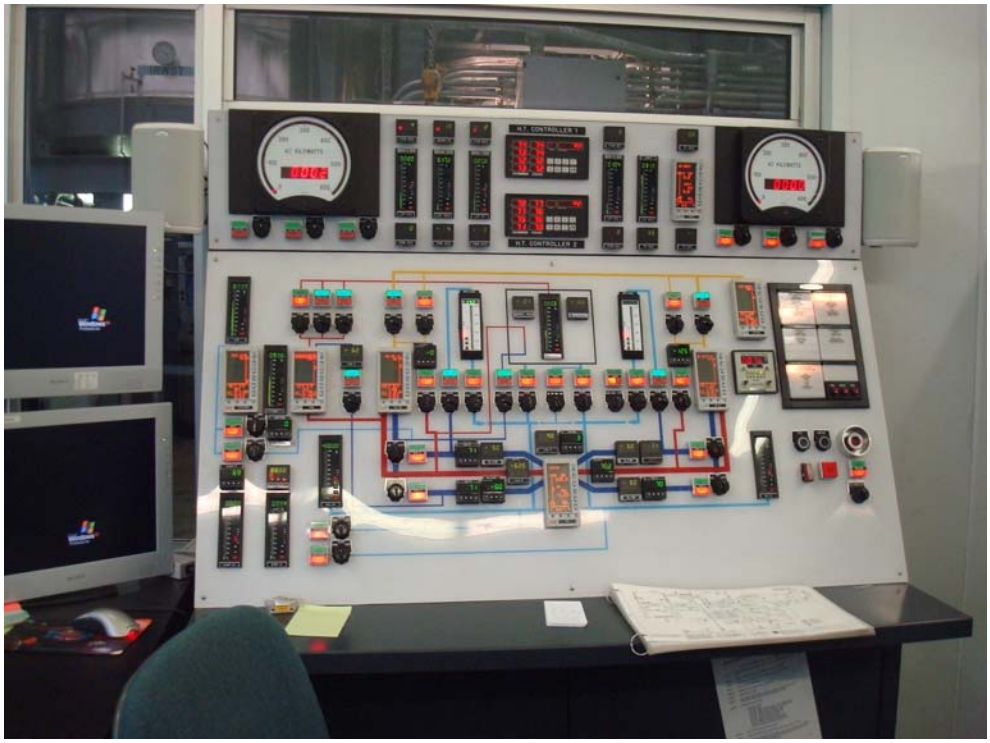

Figure 3.6. Oregon State University APEX control room.

3.2 Advanced Control Rooms and the NuScale Power Design

NuScale Power is currently designing an "advanced" nuclear power plant, to be submitted to the U.S. Nuclear Regulatory Commission (NRC) for certification. In establishing the technical basis for staffing exemption requests, NUREG/CR-6838 provides an overview of the technologies foreseen for next-generation or "advanced" nuclear power plants, and the corresponding human factors considerations (Plott et al., 2004). New plant designs may include the control of "multiple modular reactors" from a single control room, while at the same time reducing staffing levels per unit. Passive safety features, relying on natural processes such as gravity and convection instead of pumps and valves, may give the crew more time to respond to incidents. Also, technological advances in automation and HSI design may simplify operations and result in departures from traditional operator roles.

In the nuclear power industry, the design and validation of a novel system should include an operating experience review (OER) of related technologies or operating concepts in existing plants. That is, "lessons learned" should serve as input into new designs. In cases of significant departures from previous designs or operational concepts, however, this may be impossible. When relevant OER is limited, NUREG/CR-6838 recommends a review of other related industries (Plott et al., 2004). For the human factors issues faced by next generation, multi-modular reactor plants, we identified Unmanned Aerial Vehicle (UAV) control as an important source of such information. In both cases, HSI designers and evaluators are particularly interested in the mental workload and situation awareness of the operator, particularly as related to the level of

automation and the ratio of units (i.e., reactor modules or unmanned aircraft) to operators (Mouloua et al., 2001). Whereas this is a relatively new issue for the nuclear power industry, researchers have been investigating the topic for UAV control for nearly a decade.

Traditionally, multiple crew members have been assigned to the operation of a single UAV, but coincident with advances in automation technology, work is being done to invert this ratio, with a single operator supervising multiple highly automated UAVs (Dixon & Wickens, 2003). Human factors studies have looked at an operator's ability to handle anywhere from one to eight UAVs simultaneously (e.g., Dixon & Wickens, 2003, and Galster et al., 2006, respectively), with several testing a four-UAV configuration (Galster et al., 2006; Liu et al., 2009; Ruff et al., 2004). Dixon and Wickens had a dedicated workstation for each UAV, while others have integrated all UAVs into a single interface distributed over two displays.

In general, the unit-to-operator ratio is not a simple question: the complexity of assigned tasks is a closely related consideration (Galster et al., 2006). Auditory alerts can aid performance during the visually demanding multi-tasked monitoring of multiple displays (Dixon & Wickens). Dixon and Wickens also found that offloading some tasks to automation may benefit performance by mitigating workload. However, it may be difficult for operators of multiple UAVs to maintain situation awareness with high levels of automation, particularly during prolonged nominal operations (Ruff et al., 2004). Although there is some evidence of understimulation with a single UAV per operator, Liu et al. found few differences overall in performance with one versus two UAVs.

Performance in this experiment dropped off significantly with four UAVs. However, Liu et al. concluded their results helped demonstrate the feasibility of assigning multiple UAVs to a single operator.

Of note for the relevance of this study, NUREG/CR-6838 explicitly states that situation awareness and mental workload should be assessed in the design of control rooms for multi-unit nuclear plants, and that operator capabilities in such a configuration have never been evaluated in a simulator (Plott et al., 2004). The experiment presented below, then, serves as a pilot study not just for the particular goals discussed above, but also as a timely first step in a potentially major transition in the industry.

At present, NuScale Power is proposing a twelve-module nuclear plant, with all reactor modules controlled from a single control room. As drawn in early conceptual sketches, a single reactor operator is responsible for the monitoring of four highly automated reactor units. The HSIs for these four units are organized into a semi-circular workstation, with each unit taking up a spatially dedicated, continuously visible (SDCV; NUREG-0700 4.2.1-4, O'Hara et al., 2002) arc of the workstation. This design approach is intended to use spatial memory to reduce the potential for confusion between modules and for mode errors, as described by Tran et al. (2007b). Altogether, the Video Display Units (VDUs) for a workstation cluster may take up to 180 degrees of visual angle horizontally, implying that the operator will need to move around to maintain awareness of the four units, possibly on a wheeled chair.

The volume of information required to maintain "status at a glance" (NUREG-0700 section 1.1-14, O'Hara et al., 2002) for four modules implies that the design will

rely heavily on the operator's visual perception. Conversely, auditory cues (e.g., for alarms) must be used sparingly, since the HSIs for four modules are located side-by-side at a workstation, and up to twelve modules are represented in a single control room. If the system relied heavily on auditory signals, a small number of abnormal conditions could quickly lead to overload, confusion or annoyance. In cases where auditory alerts are appropriate, directional cues may be used to preattentively guide the operator to the panel for the proper module.

Whereas crew awareness and communication in a traditional single-reactor plant may be facilitated by a Group-View Display System (NUREG-0700 section 6, O'Hara et al., 2002), such a plant-wide, common display may be distracting or confusing in a NuScale plant. Each operator is assigned a set of modules and needs to monitor these. Awareness of plant-wide status, of high-level reactor health for all twelve modules, and of systems common to multiple modules is perhaps most appropriate for the control room supervisor or another dedicated crew position, who can relay information to individual operators as needed.

Minimal movement about the control room is expected; in fact, the operator will likely be confined to his or her individual workstation area to ensure a qualified person is always present at the controls. The high level of process automation will limit the need for operator intervention; that is, the primary task is typically system health monitoring. Furthermore, if the data needed to maintain situation awareness can be displayed simultaneously, there will be little need for operators to navigate the display system under normal conditions. However, one possible way to engage the operators could be requiring

them to periodically perform a tour of the module subsystems at a lower level of detail than is typically displayed.

Some discussion is anticipated between crew members in the control room, but since each operator is assigned to his or her own modules, the amount of task-related speech will be relatively low under normal conditions. Upon either an initiated or unanticipated system event, the designated operator would discuss the situation with the supervisor according to established procedures. Since the frequency of such events is much higher in the simulator, the average amount of speaking per unit time during system evaluation will likely be substantially higher than in actual operations. This has implications for the adequacy of physiological measures which are hindered by speech.

For economic reasons, a twelve-hour shift is common in traditional plants. Operators frequently move about the control room, and may monitor each other for signs of fatigue or performance decrement. With the NuScale design, the operators are assigned a more passive, supervisory role, such that the vigilance decrement, information overload, situation awareness and acceptable workload levels are key human factors considerations.

Rather than relying on the operator to understand the full implications of some failure of the automated system and to fix the problem on-line, NuScale procedures will likely prompt the operator to shutdown (i.e., “trip” or “scram”) the unit after a few brief investigative steps. This approach is more acceptable with multiple small modules than with a traditional plant, in which tripping the reactor is extremely costly in terms of plant productivity and the impact on the electrical grid. One benefit of NuScale’s philosophy of

a lower trip threshold is that mental workload and stress under abnormal conditions are not anticipated to reach the same levels as in a traditional plant, leveling the rare but sudden shifts "from boredom to terror" (Hancock and Szalma, 2003, p. 13) characteristic of traditional nuclear plant operation. Therefore, online physiological monitoring of operator state is believed to have a high cost-to-benefit ratio and is not currently being considered. This decision could be subject to change, depending on initial workload measurements obtained in the simulator. However, NuScale HFE would prefer to address any situations where stress levels exceed the operator's ability to cope by user interface and task re-design, rather than trying to dynamically tune the system during operations based on observed workload. Although adaptive automation is one way of compensating for operator under- and overload, it potentially introduces uncertainty to procedure development, training and operations, as well as to probabilistic risk analysis. It also creates another dynamic, potentially complex subsystem for which high situation awareness is essential: the operator must know his or her current task load at all times (Endsley, 1996).

To meet NuScale Power's need for data supporting an exemption request from the staffing requirements of 10 CFR 50.54(m) ("Conditions of licenses"), the relatively low time granularity of subjective measures is probably acceptable. NuScale's HFE program will need to demonstrate to the NRC that mental workload and situation awareness are at acceptable levels during both normal operations and exceptional conditions. Due to the short time window allowed in procedures before tripping a problematic module, stress and workload are not expected to reach unacceptably high levels for a significant amount

of time. In addition, utility requirements for Advanced Light Water Reactors now specify that passive safety features in such plants provide for 72 hours without operator intervention following an incident. This substantially eases the time pressure for correct operator responses to incidents, and accordingly, mitigates stress and workload. It is therefore believed that acceptable workload can be demonstrated via per-task measurements; the continuous nature of the data from physiological measures could prove useful in system assessment, but is perhaps not essential.

In a monitoring task with human operator verification and backup of automated processes, situation awareness is a key goal. Situation awareness is system- and task-specific; therefore experimental assessment will require insight into the particular conditions simulated, and their impact on the crew's ability to perform. Experimenters will work with the NuScale HFE team as scenarios are developed to identify metrics for situation awareness. A multi-disciplinary team will define concrete metrics for scenario performance in terms of what each crew position should perceive, understand and do during the simulation. Response times will be specified as part of the design and safety analysis.

Regarding further demands on the operator, there may be on the order of ten important variables to monitor frequently per module, potentially more under certain conditions. The operator may have access to many more system parameters, but this is to be determined as part of the HSI design process. At a high level, awareness of reactor health can be summed up by power, pressure, temperature and coolant level. Therefore

the awareness of present values and trends for these system parameters is especially crucial.

This is one respect in which, from a human factors perspective, nuclear power plant operation differs from other domains, such as air traffic control (ATC). In ATC the controller is assigned a working set of aircraft, and these are formally passed off over time. In a nuclear plant, an operator must likewise maintain a model of each assigned module in memory. However, he or she must also remember and understand the significant events of the recent past (i.e., on the order of hours, or perhaps the past day) per module and, accordingly, be able to detect minor changes which impact the forecasting of future system state. The latter fits with Level 3 of Endsley's (1988) situation awareness model. In the NuScale plant, exploring both past and future events will be supported by the user interface, via trending displays (e.g., NUREG-0700 section 5.1-10, O'Hara et al., 2002). In steady-state operation, there may be minor fluctuation of system parameters, but the module will tend to stay consistent. Therefore the level of display event-initiated operator activity will typically be very low; in other words, a vigilance task with a low probability of noteworthy events is expected (cf. Signal Detection Theory; Green & Swets, 1966).

A licensed reactor operator has invested time in initial and on-going training in the classroom and the simulator, building system knowledge and operating experience. While many procedures may be automated, it is ultimately the crew's job to ensure safe plant operation. As such, the most stressful or demanding tasks with the NuScale system may be those of verifying that the plant automation responds properly to unanticipated

faults or “transients,” as dictated by procedures or checklists, and instances where automation issues are detected. In case of the latter, there are backup manual procedures that the crew will have practiced in the simulator. Such rare, and potentially stressful, tasks should specifically be identified for mental workload and situation awareness assessment, to ensure adequate support from the HSI design.

Naturally, NuScale HFE desires a mental workload and situation awareness assessment methodology which: 1) is compatible with their high-fidelity control room simulator, 2) provides meaningful, readily applicable, reportable results, and 3) is cost effective. The third point notes the assumption that there may be trade-offs between the benefits of certain measures and the hurdles to their implementation, in terms of time, budget, equipment and staff expertise. Rather than identifying a rough dollar estimate at the outset with little prior knowledge of the candidate measures or their approximate costs, it seemed prudent to NuScale management and this research team to ultimately establish recommendations for several dollar amounts. We therefore aimed to recommend suites of mental workload and situation awareness measures at three levels: a small budget approach which is easily administered yet meets NuScale’s minimal requirements; an intermediate, reasonably affordable solution with additional desirable features, such as a broader base for conclusions; and the most highly recommended route, given that schedule and cost should be no consideration.

4. EVALUATION CRITERIA FOR THE SELECTION OF EXPERIMENTAL MEASURES

Chapter 2 has demonstrated the great diversity of methods proposed and used for the assessment of both mental workload and situation awareness. In discussing the state of the art in human factors for nuclear power plant control rooms, Theureau (2000) found this eclectic range of methods undesirable. While on the one hand the evaluator is provided with a wealth of measures from which to choose, workload or situation awareness comparison across systems and domains is hindered by the fact that studies use different techniques. Theureau therefore hoped that researchers would eventually reach consensus on the best measures. Sirevaag et al. (1993) took a conflicting stance, stating that early efforts in the field to identify the best single measure of workload have been abandoned in favor of using a "battery of measures" (p. 1112) within an experiment. Unless and until consensus is reached, each program or study must identify the measure(s) most appropriate for the circumstances.

Not only do measures differ between studies, but the criteria for guiding the method selection vary as well. Section 2.4.1 reviewed several sets of overlapping criteria. In this section, the various criteria are considered more closely, in the context of the goals and needs of the present effort. This analysis has resulted in seven criteria for the selection of mental workload and situation awareness measures in this study. In addition, a differential weighting system is introduced for the various criteria, for the sake of more systematically identifying the most appropriate measures and eliminating from consideration the least compatible ones.

It may be difficult to accurately judge a measure based on each criterion without significant practical experience with the measure. Therefore, in some cases, related criteria were combined into overarching, composite criteria, providing a broader base for each judgment. These criteria can be applied to both mental workload and situation awareness metrics, although certain factors may only apply to one or the other.

Table 4.1. Seven Composite Criteria in the Context of Previous Work.

HUMAN-SYSTEM PERF.	MENTAL WORKLOAD		SITUATION AWARENESS		MW/SA
IEEE Std. 845-1999	O'Donnell & Eggemeier, 1986	Zhang & Luximon, 2005	Endsley, 1995	Salmon et al., 2009	Present study
Acceptability (consensus in field)					(5) Level of adoption/ Consensus
Accuracy					(1)
Applicability					(4)
Bias					-
Intrusiveness	Intrusiveness	Intrusiveness	"does not subst. alter the construct" (p. 66)		(3) Unobtrusive-ness/ Operator acceptance
Precision					(1)
Reliability		Repeatability	Reliability	Relia-bility	(2)
Resources (required for implementation)	Implementation requirements	Convenience (implemenation)			(6) Convenience/ Cost to implement
Sensitivity	Sensitivity	Sensitivity	Sensitivity		(1) Sensitivity/ Accuracy/ Precision
Validity		Validity	Validity	Validity	(2) Validity/ Repeatability/ Reliability/ Selectivity
	Diagnosticity	Diagnosticity	Diagnosticity		(4) Interpretability/ Diagnosticity/ Applicability/ Redlines
	Operator acceptance				(3)
		Selectivity	"is not a reflection of other processes" (p. 66)		(2)
					(7) Compatibility with Problem Constraints

4.1 Sensitivity

As perhaps the most essential criterion, and as commonly found in the literature, a method must be able to distinguish between varying levels of that construct which it is intended to measure. Measures of workload, for example, which have not been proven to meaningfully distinguish between varying levels of mental demand, as correlated with varied task complexity, can be rejected. Also, a measure with finer granularity in distinguishing between levels is generally more desirable. Therefore the precision, or range of possible values, should be considered (IEEE Std 845-1999, 1999). For example, the multivariate EEG model of workload assessment presented by Berka et al. (2005) classifies workload dynamically into one of five levels, whereas NASA-TLX produces values on a continuum from 0 to 100 (Hart & Staveland, 1988). While Berka's five levels may provide greater interpretability (see discussion below) than the NASA-TLX instrument, NASA-TLX is superior in terms of precision. For subjective measures in particular, it should be noted that precision is limited both by the range of scale presented to the operator, and by the human's ability to meaningfully distinguish between very similar conditions (Hart & Staveland, 1988). Accuracy, in terms of the potential for measurement error, may be particularly hard to judge with subjective measures, and is even a complex consideration for physiological measures.

4.2 Validity

A mental workload or situation awareness metric shall have been "demonstrated to measure what it is intended to measure" (IEEE Std. 845-1999, 1999, p. 3). As a counterexample, post-experiment assessment of situation awareness may lack validity, as

this technique is reliant on memory (Endsley, 2000). While validity in general is a simple concept, the judgment of a certain measure may take many forms (Zhang & Luximon, 2005). Face validity indicates the extent to which a measure seems reasonable (Salmon et al., 2009). Correlating the data from a measure with observed performance under various conditions of controlled difficulty is one approach to investigating concurrent validity (Zhang & Luximon, 2005). Unfortunately, this technique suffers from the fact that performance and at least subjective, if not physiological, measures of workload may dissociate under various circumstances (e.g., Fournier et al., 1999; Vidulich & Wickens, 1986; Yeh & Wickens, 1988). Convergent validity is obtained by showing that various measures tend to react similarly to experimentally varied levels of task demand, thereby providing evidence for each other as valid measures (Zhang & Luximon, 2005).

Selectivity and sensitivity are related to validity, in that a measure should vary with purported changes in workload or situation awareness, but should not be affected by other factors. While some workload definitions exclude physical and emotional factors, others include these (Gaillard, 1993; Rubio et al., 2004). For our purposes, the distinction may be of relatively little importance. Repeatability (Zhang & Luximon, 2005) and reliability (IEEE Std. 845-1999, 1999) are also components of the overall validity criterion, as the data produced by a certain measure should be consistent and comparable over time and across trials. A measure that may decrease in effectiveness over repeated application, such as workload evaluation based on P300, is deemed to have inferior repeatability.

4.3 Unobtrusiveness and Operator Acceptance

For the purpose of evaluating a real-time, critical user interface, ecological validity is important. That is, the lab or simulator environment should match that of the operational setting to the extent possible. Besides the physical environment, an important component of this effort is to maintain "suspension of disbelief," minimizing reminders to the operator of the artificial setting, and thus encouraging immersion in the task and environment. Clearly, the experimental protocol should minimize disruptions of the operator during task completion. This applies to the assessment of both mental workload and situation awareness, as pausing the simulator may provide an unrealistic break from demanding duties, and may also interrupt the operator's state of awareness or mental "flow." On the other hand, Endsley (2000) argued that the technique should not alter the construct under measurement, which is a slightly different criterion.

A measure should minimize disruption of the experiment, but it should also minimize interference with operator tasks as the experiment proceeds. The measurement protocol and equipment required should not hinder the operator's completion of assigned tasks either physically or temporally. Even if it does not directly impede operator performance, bulky equipment may lead to distraction or frustration in the subject, which could reduce the operator's willingness to comply with the prescribed measure. For the same reason, subjective assessments should be relatively simple and brief.

Finally, operator acceptance may also include issues of privacy. For example, Zjilstra (1993) analyzed urine samples for concentrations of adrenaline and noradrenaline to infer workload levels in bus drivers. Whether or not this method is sensitive to

workload demands, it is at the very least inconvenient, may potentially be objectionable to participants, and may even hinder recruitment.

It should be noted that most criteria presented in the literature are stated in the positive, such as "operator acceptance," such that fulfillment of the criterion is desirable. However, the commonly cited criterion "intrusiveness" breaks this unstated convention in identifying an undesirable trait (see Table 4.1). This study preserves the same criterion, but re-phrases the concept as "unobtrusiveness" for consistent interpretation of criteria during method selection.

4.4 Interpretability

For the purpose of this study, the criterion of "interpretability" encompasses several important considerations. The primary question for this criterion is, "assuming successful data collection, what can be concluded from the results?" This issue is partially addressed in the proceeding section on level of adoption, as the more numerous the previously reported results with a given measure, the greater the basis for comparison with new systems or evaluations using the same measure.

With the exception of redlines proposed for SWAT (Reid & Nygren, 1988), there is very little evidence for or guidance on acceptable upper or lower bounds for mental workload (or, equivalently, lower bounds for situation awareness). This problem is challenging in part because meaningful guidance likely varies somewhat with system complexity, the nature and quantity of operator tasks and the potential for and implications of operator error. However, Wickens recognizes the need for such guidelines because of the implications for interpretability, particularly in demonstrating the

adequacy of new interface designs for regulatory certification (in Grier et al., 2008). That is, guidelines with some empirical basis for the acceptable values under a given measure would enable evaluators to move from conclusions consisting of "here is what we found; how does this compare to other conditions?" to "tasks X and Y fall into the acceptable range, but task Z produces unacceptably high workload and must be addressed."

In this study, another aspect of interpretability is diagnosticity. For mental workload measures, diagnosticity is the ability to separate workload further into constituent parts. This may mean pinpointing regions or structures within the brain which are taxed by a particular task (e.g., Fournier et al., 1999), linking workload demands over time to particular system events (e.g., Berka et al., 2005), or decomposing workload into demands on "the capabilities of the human operator" (Luximon & Goonetilleke, 2001, p. 230). In the context of this study, the brain topography of workload is deemed of relatively low priority, as this topic adds considerable complexity and, with current understanding, limited application. However, a basic consideration for the various types of mental resources and their limitations, such as the models proposed by Wickens (multiple resources model; 2002a, 2008) and Baddeley and Hitch (working memory; 1974, in Proctor & Van Zandt, 2008), is important during design and evaluation to avoid various types of overload. Ultimately, diagnosticity for our purposes is most useful inasmuch as a measure indicates the aspects of task demands which are problematic.

4.5 Level of Adoption and Consensus

IEEE Std. 845-1999 (1999) defines the criterion of acceptability as "the degree to which evaluators at all levels agree on the use of the measure" (p. 3). Note that

acceptability in the sense of consensus in the field is not the same as operator acceptance of a measure. The broad range of existing measures suggests a relatively low level of consensus at present, but some measures are clearly more prominent than others. For this study, consensus is therefore gauged by frequency of use in the literature. There is also an aspect of time, as newly proposed measures are unlikely to already see wide adoption. On the other hand, measures proposed early and mostly cited in previous decades may have relatively low adoption at present.

It is prudent to select measures with widespread adoption for three major reasons. First of all, relative popularity of a measure indicates that other researchers have come to similar conclusions in weighting candidate techniques, based frequently on previous first-hand experience. Secondly, in reporting results and comparing numbers between tasks and systems, it is essential to "speak the same language." For example, there is currently no way to translate some physiological measure of task workload levels with system A into a number on the SWAT continuum for comparison with previously reported results from system B. At least if multiple studies use the same measure, the resulting data can be compared at a high level, regardless of task or system similarity.

Because there are many components to interface evaluation besides mental workload and situation awareness assessment, evaluators may not be able to proclaim one system "better" or "safer" than another based on these alone. However, it may be possible to conclude that task set / user interface A is "in the ballpark of," or generally outperforms, task set / interface B, if the evaluation measures are comparable. In other words, selecting a technique more frequently used may boost the interpretability of the

results. Finally, disagreement among experts on the acceptability of a measure implies risk. That is, adopting a lesser known or more questioned measure may set a study up for critical or skeptical review of the results. Because of this risk, a measure with low adoption must provide very compelling reasons for selection, in terms of the other criteria discussed.

4.6 Convenience and Cost to Implement

Assessment of mental workload or situation awareness must take place within real world constraints such as schedule, cost and personnel resources. The practicality of a measure in these terms must be determined on a case-by-case basis, obviously, such as the availability of costly equipment or expert administrators (IEEE Std. 845-1999, 1999). However, the various methods can be generally ranked in terms of such resource demands. Zhang and Luximon (2005) noted the costs of learning and training as additional considerations of method "convenience." Equipment which requires calibration to the individual bears a cost in time to both the experiment administrator and the participant.

4.7 Compatibility with the Problem Constraints

Each of the selection criteria is to some extent colored by the needs of a particular program or the particular questions of a study. For example, interpretability should be applied to the judgment of measures in terms of the goals of the experiment. Similarly, implementation costs will be greatly reduced by choosing a technique for which the evaluators have previous experience and all of the necessary equipment. However, the

final criterion of "compatibility" is meant to address additional requirements or problem constraints of this specific study. It is suggested that the other six could be applied to new situations directly, with some problem-specific interpretation, while the seventh serves as a placeholder for other driving factors.

Judgment of measures based on this criterion requires an in-depth understanding of the tasks, the environment, the user population, and the existence of any industry-specific standards, regulations or regulatory guidance. For example, in the nuclear industry, NUREG/CR-6838 calls out certain mental workload and situation awareness measures without explicitly endorsing any of these (Plott et al., 2004). The measures identified have a precedent in regulatory guidance and are thus less risky than others. Similarly, the measures commonly used in recent nuclear applications are appealing, both in terms of level of adoption within the industry, and the potential for comparison with other systems (Ha et al., 2007). These appear to include NASA-TLX (see, e.g., Ha et al., 2007) and SACRI (see, e.g., Plott et al., 2004).

For workload in particular, there seems to be potential for underload under normal, steady-state operations. Overload appears less likely, but remains a possibility under abnormal operating conditions. In terms of the physical configuration, the operator is expected to be seated in front of a large number of digital display screens, with relatively few controls available. A common operator role will be monitoring highly automated processes.

Although precision was previously discussed in terms of the range of possible measurable values, time is another factor in comparing measures (e.g., Tran et al.,

2007a). As mental workload and situation awareness are dynamic concepts, fluctuating with task demands and system state, measures that reflect these changes over time would be the most accurate. Physiological measures provide the advantage of relatively high-resolution, continuous measurement (Tran et al., 2007a), whereas subjective measures are limited to the frequency at which evaluators choose to administer the survey, generally once per task, or possibly for each key phase within a scenario. Those measures which are already by definition subjective seem to suffer further in terms of face validity, because the participant must collapse the dynamic levels of workload or awareness over the course of multiple events into a single overall estimate, potentially hiding minimum, maximum and variability information. Subjective measures therefore present a trade-off in terms of intrusiveness and time resolution. On the other hand, if the research goals call for an overall estimate of workload for each condition considered, physiological data must be processed for reduction to such a summary form anyway.

One current direction in the field is toward real-time assessment of workload during operations for the sake of dynamic task allocation within a team, adaptive automation or adaptive user interface complexity (Berka et al., 2004; Berka et al., 2005; Hwang et al., 2008; Tran et al., 2007b). In this case, the time resolution of physiological measures is essential. If this program were to consider this possibility moving forward, then it would be logical to select the workload measures most suited to real-time assessment for both operations and for simulator-based system evaluation. In this way, a wealth of experience and data could be established from both as the mutual basis for future improvements. However, operational real-time assessment is out of the scope of

this study and is not currently foreseen for the program, so the advantages of the continuous workload measures in terms of this study are relatively insignificant, compared to the other criteria identified.

Another top-level goal of this study is to be able to evaluate mental workload and situation awareness within the same experiment, an aim which is occasionally found in the literature (see, e.g., Alexander et al., 2000; Endsley, 1993; Ha et al., 2007). Therefore, the measures selected must not only be compatible with operator characteristics, tasks and the environment, but they must also be able to co-exist without interfering with each other or the progression of the experiment (Ha et al., 2007).

Furthermore, there is considerable consensus that mental workload evaluation may be aided by selecting multiple measures, particularly if these represent different categories of techniques (Baldwin, 2003; Brookings et al., 1996; Levin et al., 2006; Sirevaag et al., 1993; Tsang & Velazquez, 1996; Veltman & Gaillard, 1996; Wilson, 2002; Zhang & Luximon, 2005). Guhe et al. (2005) went so far as to say that use of a single measure "is not reliable for workload assessment" (p. 1159). As reviewed in section 2.4.3, each category of workload measure has particular strengths and weaknesses. Therefore in terms of robustness of workload measurement and in breadth and depth of interpretability of the results it seems wise to select measures from more than one category. Brookings et al. agreed that this approach fosters a "comprehensive perspective on controller workload" (1996).

The selections must be compatible with each other, but they also must be complementary. For example, selecting two subjective measures of workload makes little

sense, as these can be predicted to provide similar information, and may even interfere with each other by confusing the subject, causing him or her to think about workload outside the formulation of the individual survey, or simply causing extended pauses in the experiment in which operational workload and stress are not accurately simulated. Similarly, relying solely on multiple cardiopulmonary measures of workload may limit the types of conclusions that can be drawn during data analysis, and may decrease the margin of error for achieving statistically significant results, as these measures are understood to be inter-related (e.g., Veltman & Gaillard, 1996, found HRV to be confounded by respiration changes). However, selection of a subjective technique, one or more physiological techniques, and task-specific performance measures would seem to provide a broad base for meaningful results. Similarly, Salmon et al. (2009) have recently suggested a "battery of different but compatible measures" (p. 499) for assessing situation awareness from multiple perspectives.

Not only are multiple measures preferred, but the identified methodology needs to handle experiments with multiple simultaneous participants (i.e., a crew; see, e.g., Patrick et al., 2006). The widely accepted approach of iterative user interface design and evaluation, with increasing interface complexity and fidelity, implies that early evaluations will likely consist of a single operator with a partial system, but that assessment with a full crew complement and realistic simulated system will eventually be necessary.

Application of subjective surveys to a crew is not much different than with a single operator. However, techniques requiring more than a pen and paper may present

time or cost scaling issues. For example, in applying physiological measures to assess events which are demanding upon the entire control room staff, there must either be a sufficient quantity of the potentially costly equipment to monitor the entire crew simultaneously, or else the simulation must be run multiple times, rotating the "instrumented" staff position. Because repeated application of a scenario with the same crew is undesirable for carryover and validity reasons, this latter approach would actually require multiple crews, which is also costly in terms of time and personnel. Team-based situation awareness assessment may also be challenging, because various staff positions should have awareness of different subsystems and at varying levels of detail. Therefore, a single uniform measure administered to each crew member may not be appropriate.

Considerations for favoring compatibility with the case-specific problem constraints are summarized here:

- Favor measures compatible with, and in some cases, identified by any federal regulations, regulatory guidance, or industry standards
- Favor measures compatible with the simulator environment
- Favor measures compatible with operator characteristics
- Favor measures which support both individual and crew assessment
- Select multiple, compatible, complementary workload measures
- Select multiple, compatible, complementary situation awareness measures
- Select compatible workload and situation awareness measures

In summary, six general criteria have been selected for this study and interpreted in terms of the needs of this study, based on previous work. In addition, an additional adaptable criterion was discussed, which is intended to capture the custom considerations based on our particular problem constraints. The values included under this custom criterion are listed below. Although the identification of these individual criteria is an essential step in making an informed, systematic selection of measures, we find that in practice there are interconnections and overlap between the criteria. However, the ultimate intent is that whatever the terminology, the proper values are identified and considered in depth throughout the selection process.

4.8 Weighting of Selection Criteria

Having identified seven criteria for the selection of assessment measures to be used in this study, we proceed to rank them, recognizing that some may be more essential than others. An absolute ranking seems unnecessary in this case, so we categorize the criteria into three separate groups: essential, important, and desirable. If a measure is deemed to not meet an essential criterion, the measure can be rejected outright. Measures which violate important criteria should only be considered if they outperform on other criteria, presenting a trade-off situation. Any identified criteria that are less than desirable did not make the original list; thus, desirable is the lowest category, which can potentially break a tie between two measures, all else being equal.

Another way to view the ranking categories is as defining thresholds for a measure to meet a given criterion. A prospective measure must meet all essential criteria,

should meet the important criteria, and ideally meets the desirable criteria. The ranking of our criteria in these terms follows.

Table 4.2. Prioritized Ranking of the Criteria for Measure Selection.

ESSENTIAL	Sensitivity, validity
IMPORTANT	Unobtrusiveness, compatibility with problem constraints, level of adoption, interpretability
DESIRABLE	Convenience

First and foremost, a measure must be reasonably expected to provide meaningful, trustworthy results, whatever the industry, problem domain or experimental goals. Assuming this is the case, it is also important to ensure the selected measure is widely recognized, will interfere minimally with the evaluation, and addresses the particular constraints and needs of the area under investigation, and to consider what can be concluded from the results. In practice, cost is also an important, driving factor, but in the absence of hard guidance along these lines, and in the interest of providing solid recommendations for the establishment of a long-term program, convenience takes secondary priority to the other criteria. For example, initially costly equipment and expertise acquisition would be amortized over the life-time of a program and may be the best approach, even if other, simpler methods are less costly at the outset.

5. SELECTION OF EXPERIMENTAL MEASURES

5.1 Analysis of Adoption of Subjective Mental Workload Measures

Level of adoption is one of our criteria for the selection of experimental measures. In December 2009 we analyzed the frequency of citation of the various subjective measures of mental workload under consideration. While citation is not equivalent to adoption, the numbers support a general notion of the relative popularity of the measures. In addition to the citation counts, we also considered the opinions of two other sources, pertaining to state-of-the-art status for the nuclear domain (Plott et al., 2004) and popularity (Proctor & Van Zandt, 2008). The results are summarized in Table 5.1.

Table 5.1. Citation Frequency of Subjective Mental Workload Measures.

Subjective Mental Workload Measure	Year of Seminal Paper	Citations on Google Scholar (as of 12/17/09)	Search Hits* in Academic Search Premier (EBSCOhost) for Last Names of Leading Two Authors and "workload" (as of 12/18/09)	Search Hits* in Academic Search Premier for Two Authors, "workload" and name of measure (as presented in first column)	Noted in NUREG/CR-6838 Appendix B (Plott et al., 2004)	Listed by (Proctor & Van Zandt, 2008) as "four of the most popular" (p. 254)
NASA-TLX	1988	172	137	125	X	X
SWAT	1981, 1988**	70, 187	4, 36	2, 20	X	X
RSME	1993	151	66	13		
Modified Cooper-Harper***	1983	113	13	9		X
Workload Profile	1996	34	3	2		X
Overall Workload	1987	27	27	3	X	
Simplified SWAT	2001	28	5****	1****		
Multiple Resources Questionnaire	2001	17	1	1	X	

NOTES:

* Includes many of the major academic sources for our literature review, including: ACM, Applied Ergonomics, Biological Psychology, Ergonomics, Human Factors, IEEE, Int'l Journal of Human-Computer Interaction, Int'l Journal of Human-Computer Studies, Safety Science. Determining total citations in this database is not straight-forward, as typos and different citation styles result in multiple sets of citations for the same paper. The results with the third approach are somewhat less reliable, because we had to choose between, e.g., SWAT and Subjective Workload Assessment Technique.

** Both (Reid et al., 1981) and (Reid & Nygren, 1988) are cited for SWAT. The latter is more heavily referenced, and is used in Figure 5.1 below.

***Author Wierwille is misspelled in the Google Scholar citation as "Wierwile."

**** This includes the paper itself. The paper was not excluded because it is not known if the hits for the other measures included their seminal papers.

The results from Table 5.1 are presented in graphical form in Figure 5.1 below to visually emphasize the citation patterns observed.

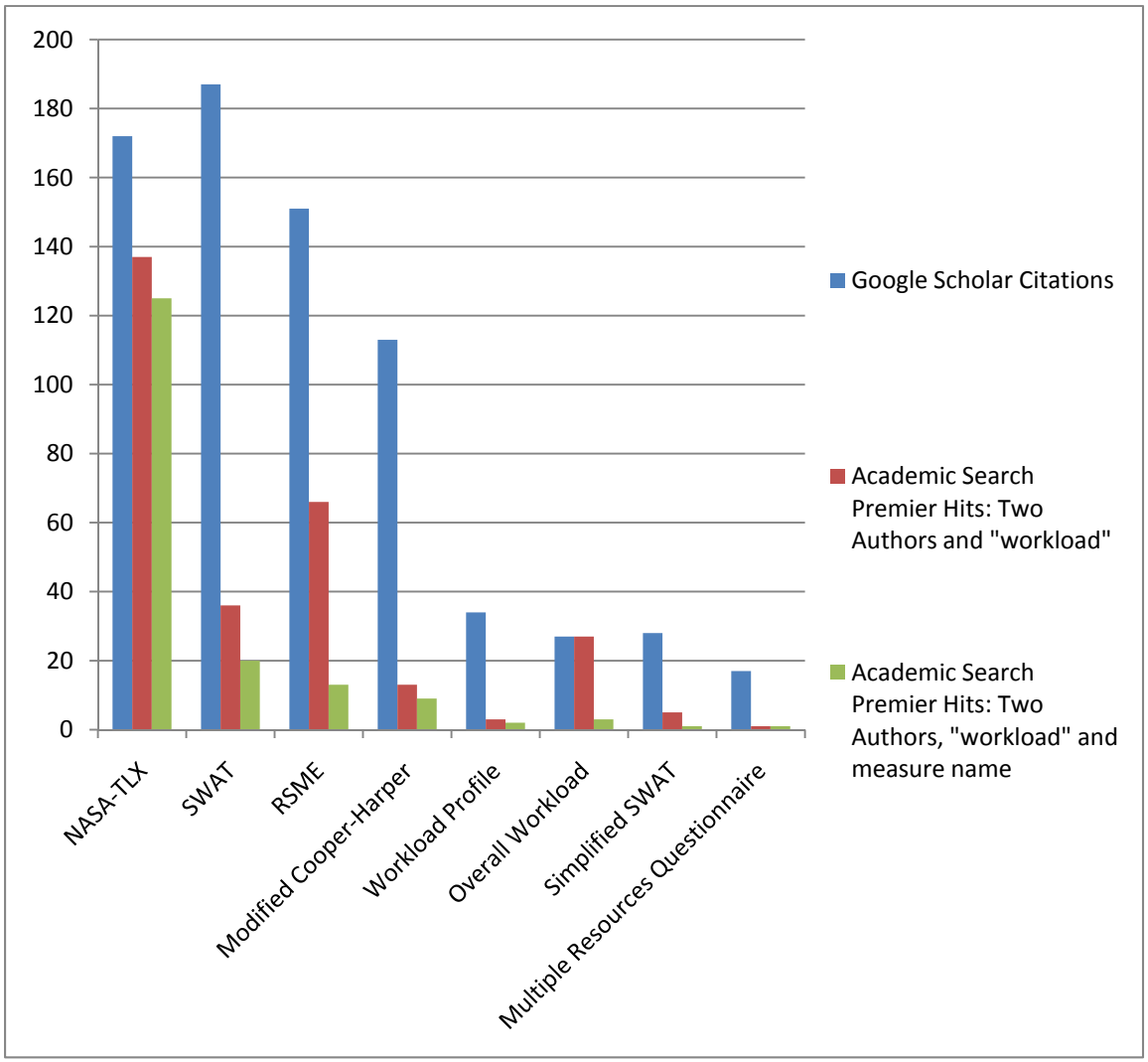

Figure 5.1. Citation Counts of Various Subjective Measures of Mental Workload (as of December 2009).

In our literature review, NASA-TLX appeared to be the most frequently applied subjective workload measure, which fits reasonably well with the trends observed above.

One surprise in the results of our citation analysis was the fact that Zijlstra's (1993) PhD dissertation, presenting RSME, was so heavily cited, considering that we encountered relatively few human factors studies using this scale. As RSME is only one outcome of this work, it is possible that the dissertation has been cited for other reasons. However, this may be true of the other measures, as well. During our literature review, we got the impression that RSME is more widely used for human factors efforts in European countries than in the United States at present. We observed with SWAT that there may be more than one single reference cited for a particular measure, which means it is possible that studies used other measures without citing the seminal work we used for our analysis. However, we believe these trends provide a reasonable indicator.

During the analysis, we recognized that it is possible that some measures may have lower citation counts because they are relatively new; these should not necessarily be rejected based on the citation analysis. To investigate this possibility, we also analyzed citation counts as a function of year of proposal, expecting that measures in existence longer should have more citations (see Figure 5.2). As expected, Workload Profile, Simplified-SWAT and Multiple Resources Questionnaire may not have been extensively cited because they are more recent. However, there is potential risk in assuming one of these will become widely adopted. The two measures which stand out as lacking citations, based on their age, are Modified Cooper-Harper and Overall Workload.

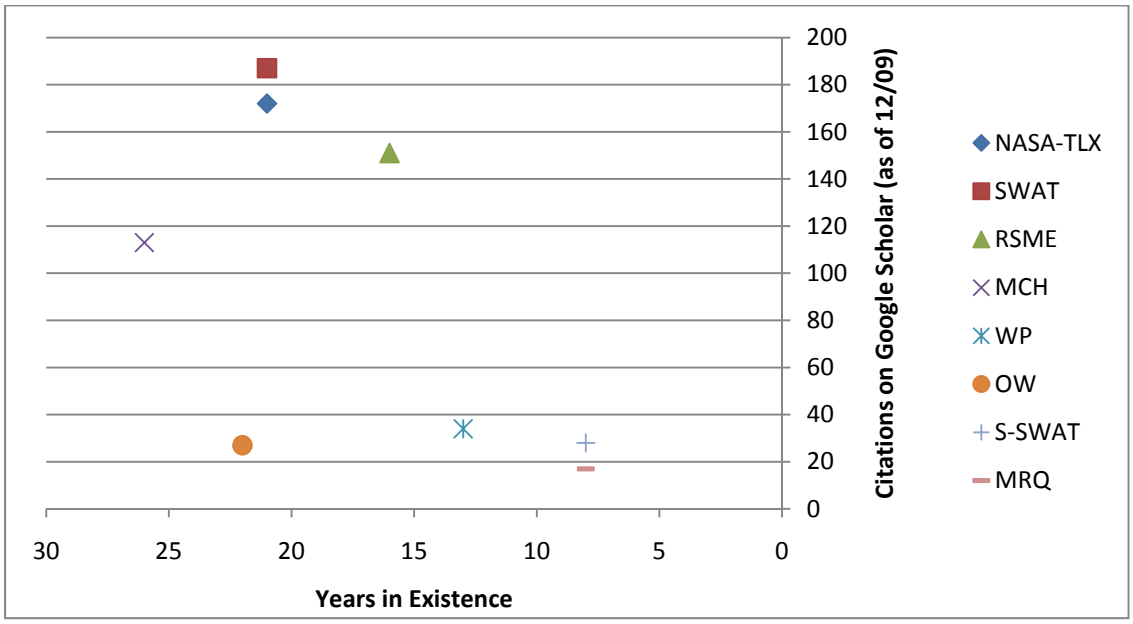

Figure 5.2. Number of Google Scholar Citations and Years in Existence for Various Subjective Mental Workload Measures (as of December 2009).

For this program's needs, we desire a measure that is widely adopted and is noted in NUREG/CR-6838 (Plott et al., 2004). The two measures fitting these criteria are NASA-TLX and SWAT, so either of these appears to be a reasonable candidate for selection. Modified Cooper-Harper fails to meet both criteria, according to our analyses. While NASA-TLX seems to be the current de facto standard, SWAT features redlines proposed on an empirical basis. SWAT, however, is criticized for its cumbersome card sorting routine.

5.2 Observation of Physiological Measures in Practice

The physiological measures of mental workload received special attention in the literature review portion of this work for two reasons. First, we believed their objective nature was a promising alternative to the more frequently used subjective techniques.

Second, the physiological measures presented us with the greatest unknowns at the outset of this effort. Although the readings gave us a good overall impression of the practical hurdles and limitations of the physiological measures, we found it prudent, to the extent possible, to observe the equipment and methods first-hand.

Our visits to different labs helped to illustrate what was noted or mentioned in passing in the related work, and to fill in the gaps of details and issues perhaps taken for granted by experts with years of exposure to their particular methods. Finally, it was instructive to describe our goals and constraints to experts and to obtain their advice regarding the theory and practice of their respective techniques. Observing EEG, EKG and eye tracking in person gave us a stronger basis for and greater confidence in our analysis in the following sections, and thus aided us in the selection of measures.

5.2.1 EEG

In October 2009 we visited the OSU Department of Psychology's Attention and Performance Lab to observe electroencephalographic (EEG) data collection and analysis. A student volunteered to serve as a participant in an EEG-based study currently being conducted in the lab. The study was investigating a particular time-dependent component of the EEG signal in response to certain stimuli. We were able to observe the process from preparations and setup, through data collection, to data filtering and analysis.

Based on the participant's head dimensions, a particular cap was selected and applied. The lab staff applied gel to the electrodes under the cap to aid in conductivity with the scalp. The volunteer had a shaved head, but we were told that the gel works

decently through longer hair, as well. The researchers recommended using the tip of the nose or a spot behind the ear as a reference node. Electrooculographic (EOG) electrodes were also placed about one eye for detecting eye activity. The EEG signals left the cap through a long bulky set of cables. The participant reported that the cabling was the greatest annoyance, as it "made movement very difficult," and that the equipment was somewhat uncomfortable, including the gel in the hair. Altogether, setup took approximately twenty minutes, which we felt was relatively long for our purposes.

In observing the participant's actions and the recorded data in real-time, we found the EEG signals extremely sensitive to movement. When the participant clenched his face, rolled his eyes, changed gaze position or blinked, several channels of the desirable EEG signals were overcome with artifacts. Additionally, we learned that there are large individual differences in the data.

Of the approximately sixteen or twenty sensors used for recording, six channels had been chosen for data analysis based on their relevance to spatial attention (i.e., channel selection was specific to the goals of the study). Automated scripts identified trials containing artifacts. An experimenter manually double-checks the results of the processing for each trial, and the data from satisfactory trials across participants are combined into a composite curve.

Prior to the visit we favored EEG-based workload measurement for its high face validity, its superior diagnosticity in time, and reported successes of applied research efforts in the literature. However, despite some discussion of common EEG issues in the literature, we were surprised by the limitations of this technique. While there is a good

chance of getting useful results, this comes at a large up-front cost of equipment and time, learning and workarounds for the evaluation staff. Also, there seemed to be the potential for high intrusiveness and low acceptance from operators. Although P300-based techniques are standard and useful in the lab, the need for a specific stimulus and many repeated trials seems incompatible with fairly naturalistic, simulator-based assessment. Frequency domain analysis therefore appears more fitting for such applications, but even so, assessment techniques in operational monitoring tasks should support relatively free head and eye movement. Wireless caps and methods for filtering and cleaning the data, as used, for example, by Berka et al. (2004, 2005), were appealing solutions to the intrusiveness and artifacts issues, but were beyond our means for the scope of this work. All of this implies that EEG analysis in the frequency domain appears to be a high-cost, high-risk, yet potentially high-reward physiological technique for the assessment of mental workload. Time domain methods present additional practical issues and are perhaps inappropriate.

5.2.2 EKG

In January 2010 we visited Jason Penry, an instructor in the Oregon State University Department of Nutrition and Exercise Sciences, for a demonstration of electrocardiographic (EKG or ECG) measurement, and his insights into our problem. We hoped to learn about the practical aspects of the equipment as applied to our environment and tasks, as well as the analysis of the resulting data. Mr. Penry stated that in his opinion, one could make a case for R-R or P-P variability (i.e., beat-to-beat heart rate

variability) as measures of mental workload or stress. He also kindly pointed us to recent, relevant scholarly papers from his field (e.g., Nunan et al., 2009; Tharion et al., 2009).

In terms of equipment setup, multiple EKG sensors are affixed to the body via gelled pads. While twelve-lead configurations are common, Mr. Penry advised that we would likely only need a three-lead configuration, using four sensors to form a rectangle about the heart. A non-expert could learn the application process with some practice. For the particular setup we observed, there were eight feet of cable leading to the recording machine, which scrolled the real-time readings across the display screen and had print capability. Data analysis is performed by hand on a paper printout. Upon completion of the trials, the sensor pads are removed and discarded. The stickiness of the pads can cause pain during removal, particularly if they have been placed on body hair. In some cases, the skin is shaved and then abraded with steel wool prior to sensor application. Privacy may be an issue, particularly for females, due to the need to apply sensors about the heart. Another female may assist with the setup to address this issue; alternatively, with a four-lead configuration, it is possible to apply the relevant sensors up the sleeve.

During the demonstration, we found that the R-peaks (a positive spike in the P-Q-R-S-T pattern) of the EKG were always easy to identify. Mr. Penry stated that lead 2 should be particularly useful. There was some noise in the data from equipment jostling or EMG (electrical signals from muscle activity) during walking, but even when the subject jumped around vigorously, the R-peaks were still visually identifiable. The wires leading to the recording machine would probably not get in the way of seated HCI tasks, and it is possible to use wrap or a mesh shirt to minimize noise from electrode jostling.

In terms of data collection advice, Mr. Penry recommended obtaining all recordings during the same time of day, particularly for a within-subjects design with multiple sessions. Also, the relative levels of fitness matter, as individuals with greater fitness and thus larger heart muscles may exhibit greater heart rate variability. Elite endurance athletes should therefore be excluded from an HCI study using heart rate measures. For practical reasons, Mr. Penry recommended we consider a heart-rate monitor on a rubber chestband, which is relatively inexpensive and more mobile than EKG equipment. Some heart rate monitor models come with software for data analysis, including HRV calculations. This avenue is relatively convenient and cost-effective, with the potential for producing useful mental workload data.

5.2.3 Eye Tracking

In February 2010 we traveled to the University of Oregon to visit the UO Department of Computer and Information Science's Cognitive Modeling and Eye Tracking Lab, under the direction of Dr. Anthony Hornof. During the visit, we observed two different remote eye tracking systems and asked questions regarding the practical aspects of data collection and the applications of eye tracking. We also learned that head-mounted eye-trackers present some advantages, but may become uncomfortable over time.

Dr. Hornof advised us that, in general, it is relatively easy to collect eye tracking data, but there is work required to link the data in time with the user's tasks and actions, and to summarize the results. The participant's gaze, as recorded by the system, may give

an indication of what he or she is aware of (i.e., this may be a reasonable measure of situation awareness). Also, pupil diameter data should be fairly easy to obtain from an eye tracking system, but it would be wise to ask the system developers how accurate the eye model is for pupil diameter measurement. Eye blinks generally show up as invalid records in the eye tracking data; this could be validated by comparing the data to frames from high-speed video of the participant's eyes. Contemporary systems provide application programming interfaces (APIs) for tailoring the system behavior to the needs of a particular study with custom software. It may take on the order of weeks to become familiar with an API and perhaps a year to develop a full application. While simply obtaining raw gaze position and pupil diameter is easy with packaged software, linking this data to stimuli or certain conditions would likely require custom development.

Dr. Hornof emphasized that good science dictates the study of the instrument as well: applied studies using an eye tracking system should analyze not just the collected data, but also the adequacy and limitations of the equipment itself. There is a quite a bit of uncertainty in the measurement of eye tracking with current methods, and there are individual differences, but these are hard to predict. If glasses are interfering with the eye tracking, it may help to tip them vertically to adjust the reflection angle away from the sensor. He also recommended that an eye tracking lab also obtain a set of vision test equipment, to pre-screen participants for any issues which may impact the results.

5.3 Criteria-Based Scoring of Candidate Measures

Based on our literature review and first-hand observation of several candidate techniques, we have compiled two-dimensional matrices crossing the measures under

consideration with the seven composite selection criteria identified. Lacking prior experience with most of the measures, we assign a value to each measure-criterion pair ranging from 4 (“very favorable”) to 0 (“very unfavorable”) using our best judgment. A score of 2 indicates some uncertainty as to the favorability of the measure for the given criterion, either due to a lack of information or mixed impressions from the literature. In many cases, the composite nature of the criteria provides a broader basis for these judgments.

The criteria-based rating matrices follow for mental workload and situation awareness measures in Tables 5.2 and 5.3, respectively. A raw score is determined for each measure as a simple sum of the scores for each criterion. The maximum raw score is 28 (i.e., a score of 4 for each of the seven criteria). A weighted score is also calculated, based on the relative prioritization of the seven criteria previously presented in Table 4.2. The numerical weightings are rather subjective and could easily be re-distributed for the purposes of future studies. The essential criteria, sensitivity and validity, are assigned a weight of 3. Convenience was deemed a desirable criterion, and is weighted as 1. The remaining four important criteria, unobtrusiveness / operator acceptance, compatibility with the problem constraints, level of adoption / consensus, and interpretability, are assigned weights of 2. Therefore the maximum weighted score possible for a measure is 60. The numbers are meaningful insofar as they serve to identify measures with relatively high and relatively low scores. The measures with high marks are deemed appropriate and reasonable options, while those with low scores are ruled out as incompatible.

Throughout this research effort, we sought for solving the specific problem faced as well as generalizable recommendations. There is a criterion designated specifically for the needs and goals of this research effort, which would tune the results to our needs. The scores in this column could easily be customized for future dissimilar work. However, while we sought general scores for the other criteria, we found that in practice, several of the criteria are inter-twined with the tasks, environment and industry. For example, the level of adoption of SACRI in the nuclear power domain is intermediate to high, but it does not apply elsewhere. Also, intrusiveness depends in part on the nature of the task and the work environment. Therefore, the matrices presented are fairly applicable to other efforts, but future studies should refer to the literature review providing the basis for our judgments, as well as the matrices themselves, for any custom considerations. We also found it inconvenient to factor into the problem-specific category the value of choosing multiple compatible measures: a careful analysis for identifying sets of such measures would require more than a two-dimensional table.

Our criteria-based scoring of measures of mental workload and situation awareness is presented in the following two tables. Those measures achieving a weighted score of at least 40 are emphasized with green highlighting.

Table 5.2. Criteria-Based Scoring of Mental Workload Measures. 4 = Very Favorable, 2 = Uncertain/Mixed, 0 = Very Unfavorable.

	Sensitivity/ Accuracy/ Precision	Validity/ Repeatability/ Reliability/ Selectivity	Unobtrusiveness/ Operator Acceptance	Compatibility with Prob. Constraints	Level of Adoption/ Consensus	Interpretability/ Diagnosticity/ Redlines	Convenience/ Cost to Implement	Total Score (Wtd.)	Total Score (Raw)
SUBJ.									
NASA-TLX	4	3	3	4	4	2	3	50	23
SWAT	2	3	2	3	3	4	2	41	19
Simplified SWAT	3	2	3	3	1	2	3	36	17
Modified C-H	2	3	2	3	3	2	4	39	19
WP	4	3	2	3	2	3	4	45	21
RSME	2	3	3	3	2	1	3	36	17
Overall Workload	3	2	3	4	1	0	4	35	17
MRQ	2	3	2	2	1	2	3	32	15
PHYS.									
HR (IBI)	3	1	3	3	2	1	3	33	16
HRV	3	3	3	2	3	2	3	41	19
BP	2	3	1	2	2	2	2	31	14
BPV	0	1	1	2	0	0	2	11	6
Respiration	2	2	2	1	2	3	1	29	13
P300 (EEG)	3	3	1	1	3	2	0	32	13
Spec. power (EEG)	3	4	2	2	3	3	0	41	17
Eye blink freq	2	1	4	3	2	3	2	35	17
Eye blink dura.	3	3	4	3	2	2	2	42	19
Pupil diameter	3	3	4	2	3	2	1	41	18
Galvanic Skin R.	4	2	3	2	2	2	2	38	17
PERF.									
Primary task	2	2	4	2	2	3	2	36	17
Secondary (artificial)	3	3	1	1	2	3	2	34	15
Embedded task (shed)	3	4	3	3	3	3	2	47	21
WEIGHTS	**3**	**3**	**2**	**2**	**2**	**2**	**1**		

Table 5.3. Criteria-Based Scoring of Situation Awareness Measures. 4 = Very Favorable, 2 = Uncertain/Mixed, 0 = Very Unfavorable.

	Sensitivity/ Accuracy/ Precision	Validity/ Repeatability/ Reliability/ Selectivity	Unobtrusiveness/ Operator Acceptance	Compatibility with Prob. Constraints	Level of Adoption/ Consensus	Interpretability/ Diagnosticity/ Redlines	Convenience/ Cost to Implement	Total Score (Wtd.)	Total Score (Raw)
SUBJ.									
SART	2	1	4	2	2	3	4	35	18
SME	3	2	4	2	2	3	1	38	17
FREEZE AND TEST									
SAGAT	2	3	2	3	3	2	2	37	17
SACRI	3	3	2	4	2	3	2	42	19
RT PROBE									
SPAM	3	2	3	3	2	3	3	40	19
PHYS.									
Gaze	2	3	4	2	1	4	0	37	16
PERF.									
Primary	1	1	4	0	2	2	1	23	11
WEIGHTS	**3**	**3**	**2**	**2**	**2**	**2**	**1**		

Table 5.4 captures the assumptions made for the analysis regarding equipment, hardware, etc. Of particular note, Nunan et al. (2009) validated the Polar S810 heart rate monitor for recording and analyzing heart rate variability. While this model has been discontinued, Polar's RS line utilizes the same technology, so their results should also apply to the Polar RS800CX (Nunan et al., 2009).

Table 5.4. Practical Assumptions for Criterion-Based Scoring of Measures.

MEASURE	PRACTICAL ASSUMPTIONS
Subjective Mental Workload and Situation Awareness Measures	Pencil and paper
Heart rate measures	Polar RS800CX and Polar ProTrainer 5 analysis software (http://www.polarusa.com)
Blood pressure measures	Upper arm cuff
Respiration	Two transducer bands
P300	Wired EEG (e.g., http://www.neuroscan.com/synamps.cfm)
EEG spectral power	Wireless cap with analysis software available (e.g., http://www.b-alert.com)
Eye measures	Tobii X60 and Tobii Studio analysis software (http://www.tobii.com)
Galvanic Skin Response	Sensors on the foot
Primary/secondary task performance	Expert observation and/or system instrumentation
Query and Questionnaire-Based Measures of Situation Awareness	Computerized

5.4 Identification of Measures

NUREG-0711 (O'Hara et al., 2004) expects state-of-the-art assessment of mental workload and situation awareness. Based on our review, we believe the state of the art presently calls for using multiple measures together. Subjective measures of workload are sensitive, easy to use and widely adopted. These are therefore desirable for "ballpark" comparison between dissimilar systems. Physiological measures present practical challenges, but they provide an alternative, more objective perspective on workload, stress or effort. For situation awareness assessment techniques, the objective measures of situation awareness based on directly querying the user about elements of the situation seem to be the most reasonable at present. However, we believe eye gaze information is a promising technique, particularly with the growing use of remote eye tracking systems. Finally, inferring workload or situation awareness from performance can be problematic, but the redline for system acceptance is most meaningful in terms of assessing performance directly, without the intermediate inference step. In other words, performance should be viewed as an indicator of design adequacy, on par with mental workload and situation awareness, rather than as a technique for assessing these constructs. We believe that a state-of-the-art assessment of user interaction with a safety-critical system should include measures from each of these categories, providing a broader base for making decisions.

Based on the results of the analysis above, we selected a set of measures that we deemed reasonable for addressing our needs. For assessing mental workload, this

includes the subjective measure NASA-TLX and the physiological measures of pupil diameter (measured with a remote video eye tracking system), heart rate and heart rate variability (measured with a wireless heart rate monitor on an elastic chest band). For assessing situation awareness, this includes an analysis of eye gaze (measured with the eye tracking system) and the objective questionnaire SACRI, to be adapted for a control room with multiple reactor modules. Task performance should also be evaluated, but the selection of appropriate performance measures is system- and perhaps even task-specific. We assume automated collection of some aspects of user performance here. This proposed set of measures, which we believe best fits our weighted criteria, is somewhat costly, due primarily to inclusion of the remote eye tracking system.

As mentioned previously, we deemed it wise to make recommendations at three budgetary levels. The set of measures outlined above is an approach with intermediate costs for equipment and setup (approximately $40,000, assuming a single remote eye tracker, a single heart rate monitoring system, and software for computerized SACRI). In our view, a low-cost state-of-the-art solution should still contain techniques from various categories of measures. For a budget on the order of $5,000, we would recommend NASA-TLX (paper or computerized), wireless heart rate monitoring, computerized SACRI, and a subjective measure of situation awareness. For a small budget, a subject matter expert could be retained to assess performance and situation awareness. However, the cost of this expert analysis is omitted from the dollar figure above.

Given time and resources, it would be appealing to attempt to monitor EEG via a wireless headset, with automated software classification of the data. Some studies have

reported success with this approach, but the practical hurdles are numerous. For the purposes of this study, we deemed EEG-based assessment unfeasible, but we believe it is worthwhile for a top-dollar budget. Therefore, for this budget level, we recommend this technique either in addition to or in place of the heart rate monitoring, with all other measures maintained from the intermediate budget approach. The price quote we obtained for such a system is comparable to the cost of a remote eye tracker, so we estimate this large-budget approach to cost roughly $65,000. If multiple eye tracker or EEG headset systems were required, this figure would grow accordingly. Unless a reliable, turn-key, off-the-shelf system can be identified, we believe this method would also require greater expertise and preparation prior to administration.

Table 5.5. Recommended Measures at Three Budgetary Levels.

Budget Level	MW Measures	SA Measures	Task Perf. Measures
Low (≈ $5,000)	NASA-TLX Heart rate Heart rate variability	SACRI (i.e., SAGAT-based approach) Subjective measure	SME observation
Intermediate (≈ $40,000)	NASA-TLX Heart rate Heart rate variability Pupil diameter	SACRI (i.e., SAGAT-based approach) Eye gaze	Computerized data collection SME observation
High (≈ $65,000)	NASA-TLX Wireless EEG Heart rate? Heart rate variability? Pupil diameter	SACRI (i.e., SAGAT-based approach) Eye gaze	Computerized data collection SME observation

Regardless of the measures selected, there will certainly be costs for training, simulation preparation, data collection and analysis. Obviously, the costs of data analysis will scale with the number of measures used and the level of analysis (e.g., scenario averaging versus moment-by-moment investigation).

All things considered, we believe the intermediate option is a reasonable, cost-effective solution for our problem constraints. Modifications may be required for application to other domains. For example, just as SACRI was adapted from SAGAT for the nuclear power domain, a program in another industry may also need to perform a situation awareness requirements analysis to adapt the technique (see, e.g,. Salmon et al., 2009). Remote eye tracking may not be practical in task environments with a very large field of view. Also, heart rate and heart rate variability data is probably not useful for tasks requiring significant physical activity.

As the next step in this research effort, we sought to validate the measures selected at the intermediate cost level. This consisted of a pilot study with a small set of participants performing representative tasks in a lab resembling the target environment. By applying the techniques within such an experiment, we gained initial experience with the preparation, setup, data collection and analysis for each individually, as well as any unforeseen interference in practice. From this first-hand experience and the resulting data, we should be able to conclude this study with greater qualifications and confidence in our final recommendations, either confirming or revising our initial, tentative set of measures. The proceeding study is thus intended to serve as a proof of concept of the set

of techniques we have chosen. Given this aim, the emphasis is placed on general impressions and applicable findings more than statistically significant results.

6. EXPERIMENTAL DESIGN

6.1 Goals and Research Questions

A modest validation experiment was devised. The primary goal of the experiment was to put the techniques into practice, and to reflect on the results from three perspectives. First, did the various measures provide useful data? Second, how did the administrators feel about applying the techniques? Third, how did the participants feel about the demands which the various techniques placed on them? In effect, this was a pilot study for the recommended experimental methodology, with the understanding that our final recommendations would possibly need to be revised based on the findings.

A secondary purpose of the experiment was to begin an investigation of the demands, workload and situation awareness experienced by operators in a multi-modular reactor control room. Such research will be needed, and has not been undertaken to date (Plott et al., 2004). For example, how do workload and situation awareness compare for an operator assigned to one versus multiple modules? How do these constructs compare between active, procedure-driven control tasks and more passive monitoring of automated systems?

In part because of the tension between the two main purposes, this study sought a balance between ecological validity and experimental control. The research goals and the questions which we sought to answer are outlined here, and these questions are subsequently addressed in the proceeding chapters 7 and 8, Results and Discussion.

1. Goals
    a. Pilot study for the techniques selected
        i. Mental Workload: Heart rate, heart rate variability, pupil diameter, NASA-TLX
        ii. Situation Awareness: SACRI (modified for four instances of the PWR simulator available), remote tracking of gaze
        iii. Initial experience and lessons learned
            1. Data – proof of concept
            2. Experience/feedback of researchers
            3. Experience/feedback of participants
    b. Pilot study for human factors evaluation of multi-modular reactor control room
        i. Aim for reasonable level of ecological validity
        ii. Aim for reasonable level of experimental control
2. Research Questions
    a. How does the prescribed methodology work in practice?
        i. As researchers, how manageable is the methodology?
        ii. How receptive are the participants to the various measures?
        iii. To what extent do the measures interfere with task performance and scenario validity?
        iv. Are the physiological measures sensitive to variation in task demands?

v. How do the physiological measures compare in terms of interpretability?

vi. What can we learn from the continuous physiological data on different time scales (i.e., moment-to-moment vs. between conditions)?

vii. Do the physiological and subjective measures provide converging or contrasting views of workload?

viii. Are there issues in the adaptation of SACRI for multiple units?

ix. Do the simulator freezes present problems for the continuous (i.e., physiological) measures?

b. What can we learn about the human factors issues facing multi-modular reactor control and monitoring?

i. How do monitoring and control tasks compare in terms of mental workload and situation awareness?

ii. How do mental workload and situation awareness compare in various conditions?

iii. Does a supervision task with multiple modules elicit high or low mental workload?

iv. In a supervision task, how does situation awareness with multiple modules compare to one module?

6.2 Challenges and Limitations

In seeking to answer the questions above, we faced a number of challenges. Therefore trade-offs were made based on the goals of the study and the resources available. First of all, for a full-fledged nuclear power plant human factors study, extensive training to the point of nearly asymptotic operator performance is required (NUREG-0711, O'Hara et al., 2004). This was simply not possible for the scope of this study, for which a single session of three hours per participant was foreseen. Novice users are especially prone to overload (Gevins & Smith, 2003), and a lack of expertise may hinder the search for relevant information, thus limiting situation awareness (Sarter & Woods, 1991). Furthermore, as the concept of operations under study was new, even if participants had previous operational experience, some un-learning of training and habits may be necessary (see Gaillard, 1993). For potentially challenging tasks, participant motivation was also a key issue: the tasks needed to be representative for the sake of some level of ecological validity, but not overly boring or challenging to the point of discouraging the participants from aiming for good performance.

A small yet somewhat diverse set of tasks was devised, which would ideally manipulate the nature and magnitude of task demands, for the sake of validating the workload and situation awareness measures chosen. For practical reasons, each session consisted of one participant assigned to multiple reactors; crew interaction was simulated with a confederate control room supervisor. A full study would certainly require a full-scale simulator and a full crew complement, but it may make sense for such studies to

gradually build to that level of integrated human-system evaluation, as the interface design and operational concepts are iteratively refined.

The effects of long shifts and fatigue on vigilance and performance could not be adequately investigated here, again for practical reasons. Hwang et al. (2008) suggested that sessions longer than those in their study should be used to investigate the effects of low workload on performance over time; this will be an important consideration in future studies. However, we sought to include brief periods (i.e., on the order of minutes) of relative inactivity or low task demands, for investigating the effects of nominal, steady-state operation on the operator's task engagement, workload and situation awareness. That is, we needed a mix of noteworthy and potentially stressful events and undemanding or boring segments, and we needed to measure the workload and situation awareness during both. Finally, we recognized that it might be challenging to produce realistic levels of stress and the pressure to perform quickly and accurately in a usability lab-based control room simulation, even during periods of potentially unsafe system incidents. In other words, it could be difficult to achieve suspension of disbelief in a mockup simulator to the point that the psychological stress of operating a safety-critical system can be accurately simulated.

6.3 Participants

For the study, we sent a recruitment email to students in the OSU Department of Nuclear Engineering and Radiation Health Physics. We sought people with Pressurized Water Reactor (PWR) operations experience either in the U.S. Navy or a commercial nuclear power plant, and noted that previous experience with a PWR simulator was

desirable. In the event that we could not recruit a sufficient number of participants meeting these qualifications, we planned to also accept participants with graduate standing in the department and a fundamental understanding of PWR-based power generation.

Six students responded to the initial recruitment email and all six scheduled experimental sessions. One potential participant (P104) did not come at the scheduled time and did not respond to follow-up contact. Thus five enrolled and participated in the study. Each of these met at least one of the qualifications for participation. Two participants had Navy experience: one had three years of experience as a reactor operator (P106) and one had been a trainer for nuclear reactor operation for three years (P101). Two other participants (P102, P105) had attained Senior Reactor Operator status on TRIGA reactors (for academic research); one of these (P102) had also previously worked in a nuclear plant under the Tennessee Valley Authority as an assistant to the unit operator. The final participant (P103) had no previous operations or simulator experience, but was a first-year Master's student in Nuclear Engineering. Of those with operating experience, two were undergraduates in Radiation Health Physics, one was pursuing an M.S. in Nuclear Engineering, and the other was pursuing an M.S. in Radiation Health Physics. Including P103, the participants averaged 3.6 years of experience in reactor operations, on a variety of systems. The age of the participants ranged from 22 to 32 years, with a mean of 26.2 years.

6.4 Lab and Equipment Configuration

The OSU Human-Computer Interaction (HCI) Lab was the site of the experiment. The room was configured for one participant serving as operator to monitor four instances of the reactor simulator (see Figure 6.1). Four PCs running Windows XP each hosted an instance of the commercially available reactor simulator PCTRAN (http://www.microsimtech.com/pctran/). For this study, a two-loop PWR was simulated by software provided by the International Atomic Energy Agency (IAEA) to member countries for educational purposes. Each simulator PC featured two vertically stacked 19- or 20-inch flat-screen displays, with the lower monitor displaying the main mimic screen, and the upper monitor devoted to a time-based trend plot. The four PCs shared a single mouse and keyboard over the lab network via the open source software package Synergy (http://synergy2.sourceforge.net/). A red "master plant alarm" light was placed to the right of the screen array.

The participant was seated at a desk approximately 67 inches from the displays, to ensure that the eye tracking device could record gaze activity anywhere on the eight displays. For eye tracking, the Tobii X120 was placed on the participant's desk and aimed upward toward eye level. Setup included configuration of the scene in the Tobii Studio software. Small dots were placed near the four corners and in the center of the monitor array for manual eye tracker calibration (Figure 6.2). The eye tracker was set to record at 60 Hz.

During the experiment, the participant wore, under the shirt, a Polar WearLink chest band and transmitter set in R-R recording mode (i.e., it recorded every heartbeat).

The transmitter device sent this data wirelessly to the Polar RS800CX wristwatch-based training computer for recording. Heartbeat information was subsequently downloaded to the Polar ProTrainer 5 software on a PC for analysis at 1 ms accuracy. A new recording was initiated for each experimental condition (i.e., simulated scenario). Within a recording, a lap feature on the wristwatch served to mark the times when the simulator was frozen and resumed.

The control room supervisor's station included a PC for initiating the eye tracker calibration process and recording the eye data. Simultaneously, this PC also recorded digital audio and video of the session received over Ethernet from two Cisco WVC210 Wireless-G Pan Tilt Zoom Internet Video Cameras. The primary camera was located in front and to the right of the participant and captured the experiment dialogue, as well as the participant's facial expressions, posture, actions and desk surface. A secondary camera provided a bird's-eye view of the desk surface and the mimic displays, from above and behind the participant.

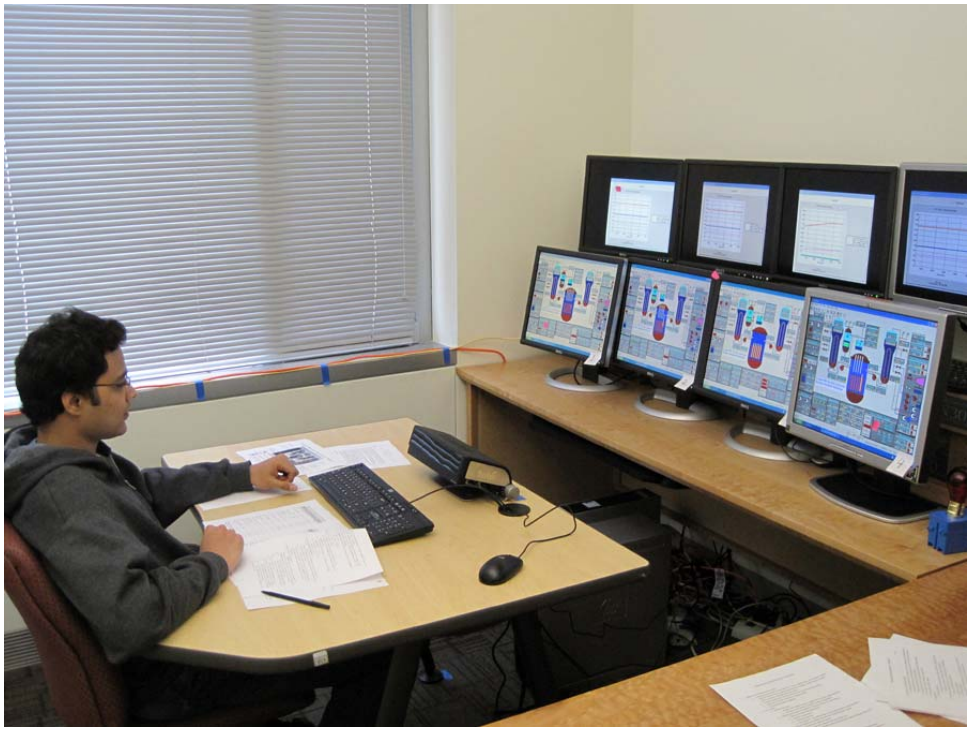

Figure 6.1. Simulated Control Room Setup.

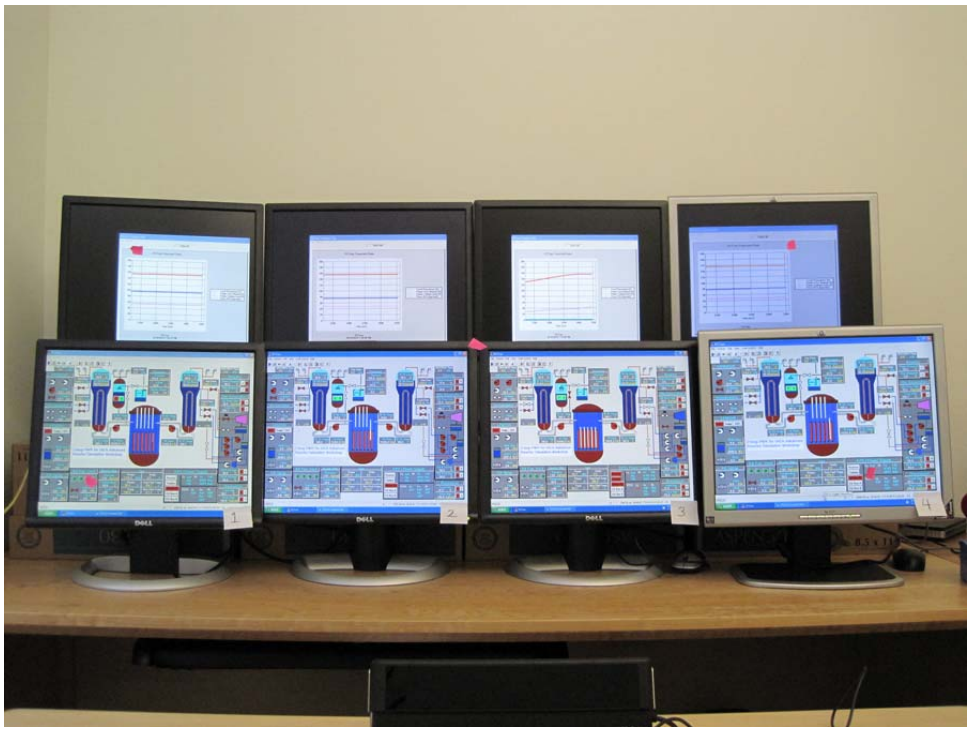

Figure 6.2. Four Reactor Monitoring Setup from Operator's Perspective.

6.5 Survey Instruments

We used NASA-TLX and an adaptation of SACRI, both in pen and paper format, for measuring mental workload and situation awareness, respectively. These were always administered together, with NASA-TLX administered first, because it takes less time and did not appear to interfere with situation awareness in pre-experiment pilot testing. The NASA-TLX subscale weighting procedure was administered immediately following the last task because participants could not be expected to meaningfully compare the various dimensions prior to experiencing the types of tasks studied. Participants were instructed to reflect back on the various tasks they had performed across conditions in order to fill out the weighting worksheet.

A subset of the questions contained in SACRI were used for this study, based on the information relevant to and displayed by the simulator. Twelve parameters were selected for inclusion in the adaptation of SACRI, some of which would have two values, one for each of plant secondary sides A and B. Tables of typical values were provided for reference for steady-state operation at 100% and 75% power, and for a temporary hold point at 60% power. These tables were necessary because of the category of SACRI questions which compares current parameter values to typical values.

We adapted SACRI to account for multiple reactors. The twelve questions for a particular freeze were randomly generated, given some constraints. First, each SACRI questionnaire was required to ask at least one question about each of core power, pressurizer pressure or pressurizer level, as these are key indicators of system health and thus were included in the trend plots. No module-parameter pair was repeated within a

questionnaire, not even with differing time references (e.g., recent past versus near future). Also, the twelve questions were ordered by module, such that all questions about reactor 1 were presented first, and so forth. With four units in operation, a minimum of two questions and a maximum of four questions were posed per unit. Questions were not ordered with respect to time, however. Rather than skipping questions for which they did not know the answer, participants were instructed to give it their best guess, as recommended by Endsley (1995). Similar to Hogg et al. (1995), the twelve SACRI questions were supplemented with questions probing the participant's high-level understanding of the system state, reactor health, and subjective awareness on a five-point scale. See Appendix A for an example.

For convenience, the simulator freeze points were pre-determined, but they nevertheless should have been relatively unpredictable to the participants. Also for convenience, one questionnaire was randomly generated for each freeze and was common to all participants. On average, there was one freeze for every 11.7 minutes of operation, with a minimum of 6 minutes of uninterrupted operation and a maximum of approximately 18 minutes. See Endsley (1995) for helpful recommendations regarding the implementation of SAGAT; these should apply to SACRI as well.

6.6 Methodology

6.6.1 Welcome and Informed Consent

Each participant reported to the HCI Lab on the OSU campus for an approximately three-hour session. Following a brief welcome, informed consent was obtained. Prior to obtaining the participant's signature, we emphasized the three

following items. First, the participant was free to end the session and leave at any time, but he or she would only receive compensation if the session was completed. Second, the participant was not under test in the experiment and should not feel personal pressure to perform. Rather, this was a pilot study looking at an experimental methodology and the operational concept of a multi-modular nuclear plant control room. Finally, there would be two built-in restroom, water and stretch breaks.

6.6.2 Instructions and Training

Instructions and training followed informed consent. The experiment administrator described the control room concept, with four highly-automated, completely independent pressurized water reactors on-site and assigned to a single control room, under the supervision of a single operator (i.e., the participant). One administrator served as the control room supervisor, who would give directions and answer questions, and a second assisted in carrying out the experiment. Each of the four reactors was presented to the operator by two vertically stacked screens, totaling eight screens, with a shared mouse and keyboard (Figure 6.3). One or more reactors may be in operation at a time, and each may be set at different desired power levels.

The operator's primary task was monitoring of the highly automated system. During steady-state operation, process parameters should remain quite stable. If the operator detected fluctuations, parameters crossing pre-defined thresholds, or system alarms, he or she should report such incidents to the supervisor and log them in an "Incident Log." Additionally, the operator should log the values of seven key parameters

per reactor unit every five minutes in a “Periodic Log.” Operating rules dictated that the operator should not take corrective action without supervisor approval, nor should he or she trip the reactor. Although the interface provided for control of valves and pumps, these controls were assigned to automation and should not be manipulated by the operator.

During the instruction phase, the supervisor started one instance of the reactor simulator to demonstrate how the plot on the upper screen should appear under normal, steady-state operation. This plot featured four key parameters indicating system health: pressurizer level, pressurizer pressure, core power percentage and reactor power percentage (i.e., turbine load or electrical output percentage). The supervisor described the basic principles of PWR power generation, the major components of the reactor, and the pertinent elements of the mimic interface on the lower screen. The supervisor also recommended a scan pattern strategy which coincided with the ordering of key parameters, and the typical values and lower and upper thresholds in a table for reference. The supervisor explained that in steady-state operation “normal” meant not just that a parameter was within these bounds, but that it was close to the typical value and not trending up or down or fluctuating. Also, visual indication of system-detected leaks (e.g., in a loss of coolant accident or a steam generator tube rupture) was described as an emergency situation which should be reported to the supervisor. A simple alarm system consisting of a red light master plant alarm and a buzzer was demonstrated. The number of times that the buzzer rang encoded the number of the problematic reactor: 1, 2, 3, or 4.

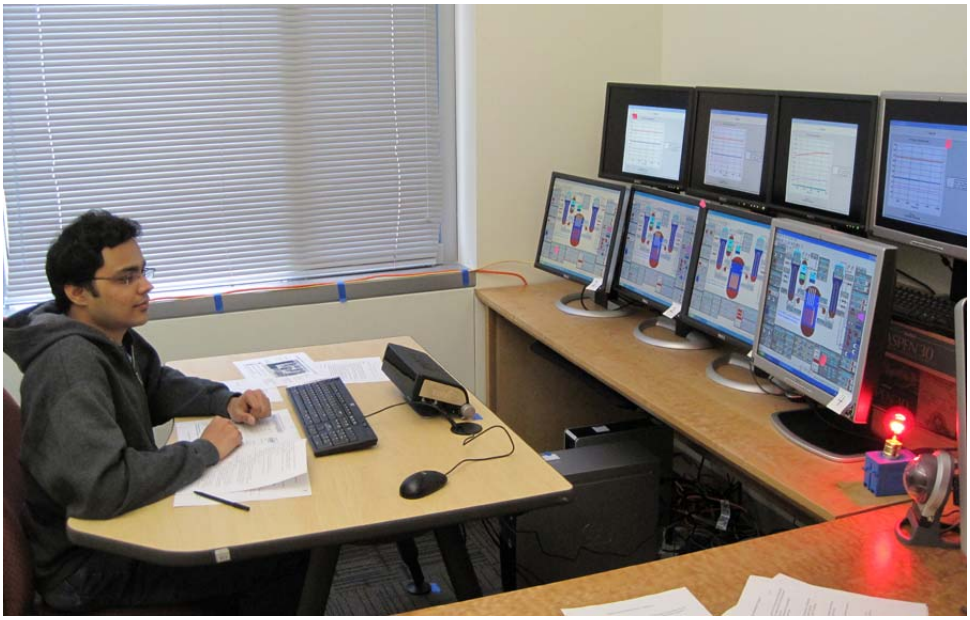

Figure 6.3. Master Plant Alarm Light.

After answering any questions, the supervisor gave the participant the opportunity to practice monitoring a single reactor for approximately five minutes, in order to become familiar with the task, the interface, the locations of the parameters to be monitored and their typical values. Next, the supervisor froze the simulator, blanked the screens, and described NASA-TLX and the modified SACRI questionnaire. For the preceding five minutes of practice, the participant filled out NASA-TLX and a brief practice version of modified SACRI to become familiar with these prior to the experiment. As this concluded the instructions and practice phase, the participant donned the heart rate monitor chest band, in privacy if he or she so chose, and was ready for data collection.

The data collection phase consisted of a five-minute resting baseline for the physiological measures, three major conditions of approximately forty minutes each (within-subjects design), and a post-session interview and debriefing. Breaks were

provided after the first and second conditions encountered. At the start of each condition, and upon resumption after each simulator freeze within the conditions, the eye tracker was manually calibrated by directing the participant to look at various dot markers on the screens. The first condition consisted of monitoring a single reactor, and was considered as a baseline task load condition and additional familiarization and practice time. Therefore, this condition was always presented first. The order of the second and third conditions was counter-balanced.

6.6.3 Resting Baseline

Similar to other studies using physiological measures, a resting baseline was obtained for comparison with the task-manipulated conditions. An administrator instructed the participant to relax and not think about anything stressful for five minutes. The participant was also asked to keep his or her gaze directed within the area of the eight display screens, for the sake of successful eye tracking. Pupil diameter samples were recorded via the eye tracker and beat-to-beat heart data was recorded via the chest band for five minutes.

6.6.4 Condition 1 – Monitoring of One Reactor

In the first condition, reactor 3 was in steady-state operation at 100% power, with the displays of the remaining three reactors blanked. The operator was assigned a monitoring and periodic logging task. At the six- and thirteen-minute marks, the simulation was frozen and the displays were blanked for administration of NASA-TLX

and then modified SACRI. Starting at the seventh minute, a slight rod insertion caused a power drop below the threshold. At approximately the tenth minute, the power began to increase again. The power continued to rise, past 100%, and gained in rate at approximately the 22nd minute due to a moderator dilution. At the 23rd minute, the master alarm was activated for unit 3. The operator was instructed to log seven key process parameters for the reactor every five minutes, starting in the fifth minute, and to report and log incidents. After 25 minutes of simulation, the scenario ended and NASA-TLX and modified SACRI were administered once more.

6.6.5 Condition 2 – Monitoring of Four Reactors

Condition 2 was designed to be similar to the first condition, except that all four reactors were displayed and were operating in steady-state at 100% power. The simulators were frozen and the displays were blanked after sixteen minutes for NASA-TLX and modified SACRI, and again after 25 minutes of simulation time at the completion of the scenario. Starting in the second minute, a load rejection on reactor four caused a significant power drop, with the reactor eventually stabilizing at 80%. At about the tenth minute, the supervisor instructed the operator to manually control this unit to return to 95% power at a rate increase of 5% per minute. Starting in the 12th minute a steam generator tube rupture occurred on reactor two, causing the pressurizer level to begin to decrease. In the 19th minute, a loss of coolant accident on this same unit caused this loss to accelerate. An alarm was generated for unit 2 in the 21st minute.

6.6.6 Condition 3 – Monitoring and Procedure-Based Control with Four Reactors

Condition 3 was fairly different from the other two conditions, in that a more active procedure-based control task was investigated. The condition began with a briefing, in which the supervisor provided the operator with three paper-based procedures and explained the goals of the scenario. Reactor 3 was at 100% power, and would be brought down to an intermediate hold point of 60% power, and then to 15% power for turbine trip and reactor shutdown. To partially compensate for the decreased power generating capability, reactors 1 and 4 would both be increased from 75% to 95% power. The supervisor instructed the operator to begin the shutdown procedure for reactor 2 immediately, and subsequently to initiate the power increase procedure for unit 1 while waiting for reactor 3 to stabilize at the 60% power hold point. Reactor 2 was operating at 75% power and would not be controlled in this scenario; the operator was asked to complete periodic monitoring log entries every five minutes for this unit only to maintain awareness. An inadvertent rod insertion incident on reactor 4 starting in the tenth minute was intended to cause the operator to decide not to proceed with control of this unit at a procedure entry decision step; naturally, this was omitted from the scenario briefing.

The procedure formatting was modeled after General Operating Instructions (GOI) 2-2 and 2-3 for start-up of Trojan Nuclear Power Plant, which is no longer in existence. Procedure content was developed with the help of subject matter experts Brandon Haugh and Ross Snuggerud of NuScale Power, both with previous commercial plant experience.

6.6.7 Interview

Upon completion of the final condition, a NASA-TLX head-to-head dimension weighting worksheet was explained and administered. A post-session, semi-structured interview lasting approximately fifteen minutes was then conducted to learn more about the participant's experiences in terms of the assigned tasks, the control room interface, the physiological measures, and the instruments. Relevant personal, training and background information was also obtained. Two main goals of the interview were to potentially help explain any irregularities in the data and to more directly obtain the participant's view of the operational concept, the incurred workload and the levels of situation awareness attained, as well as the methodology. Finally, the participant was given the chance to ask about any aspect of the experiment for a chance to de-brief, and to give any recommendations for future similar studies. The script for the interview is included in Appendix B.

6.6.8 Compensation

Upon completion of the session, participants were compensated with 50 dollars in cash. Because they came from a single department at the university, participants were asked to not discuss the experiment with others until the end of the study.

7. RESULTS

7.1 Task Performance

All participants generally did well at reporting and logging abnormal conditions. All participants also performed with comparable efficiency and accuracy in carrying out procedures. However, some maintained the periodic logs better than others (Figure 7.1). Because we did not specify prioritization for sub-tasks, the periodic logging could be viewed as either one of several time-shared primary tasks or a secondary (embedded) task.

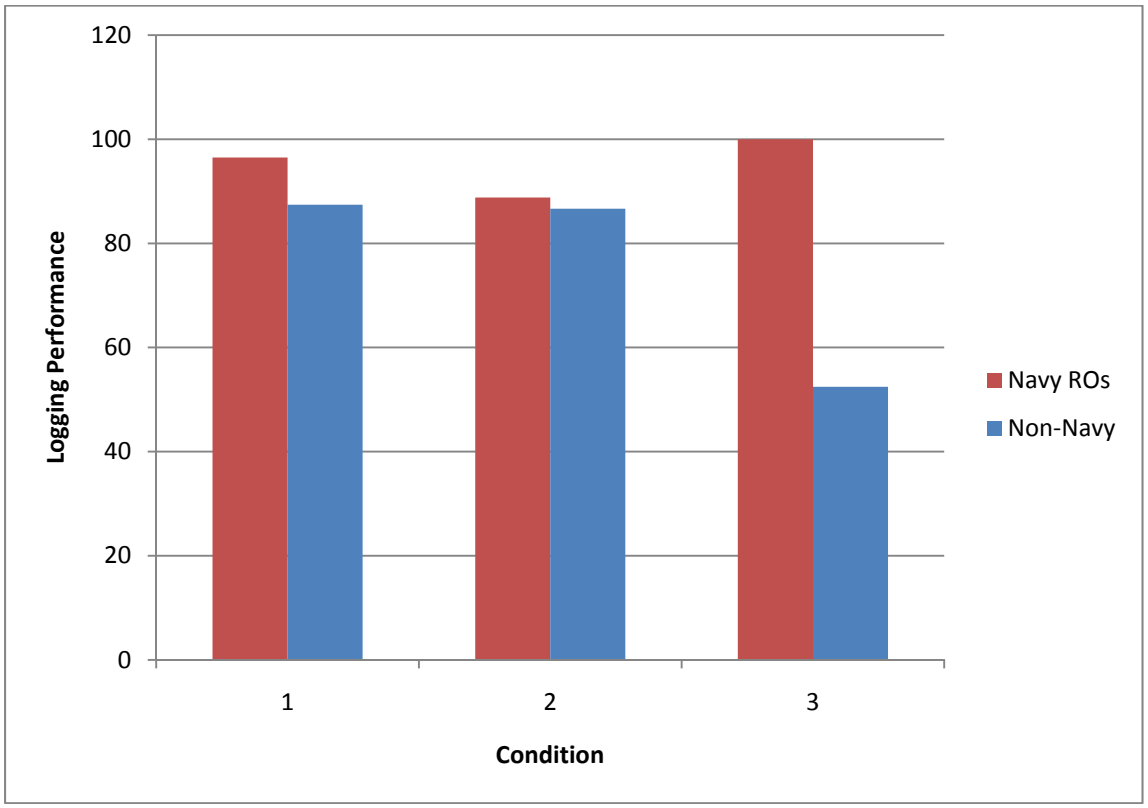

Figure 7.1. Periodic Log Performance by Condition, Separated by Operations Background.

It should be noted that the periodic logging instructions varied per condition. Therefore interpretation of the logging performance as a secondary measure of workload between conditions is not ideal. The instructions are summarized in Table 7.1.

Table 7.1. Periodic Log Instructions per Condition.

Condition	**Task**	**Log Period**	**Instructions**
1	Monitor reactor #3	5 minutes	Reactor 3 with precise timing
2	Monitor reactors #1-4	5 minutes	All four within each 5-min interval
3	Monitor reactor #2 and control #1, 3, 4	5 minutes	Reactor 2 with precise timing

Periodic logging performance was scored as follows. There were 100 possible points for each required entry. The score obtained for a log entry decreased linearly with the tardiness of the entry at a rate of 20 points per minute. With this approach, an entry appearing five minutes late received zero points and was interpreted as a prompt entry for the next period. For example, a log entry required at the tenth minute and entered at 11:30 received a score of 70. An overall score per participant per condition was calculated by dividing the total points achieved by the number of log entries expected. The results from Figure 7.1 are discussed in detail in section 8.1.

7.2 Mental Workload

7.2.1 NASA-TLX

NASA-TLX ratings were obtained three times in condition 1, and two times each during conditions 2 and 3, per participant. The weighted composite results are presented in Table 7.2. Individuals differed in the variability of their ratings between conditions (compare, e.g., P101 and P103). At the completion of the experiment, the head-to-head dimension comparison technique was used to determine the relative weights of the six NASA-TLX dimensions for the types of tasks under consideration. The composite weights are presented in Figures 7.2 and 7.3.

Table 7.2. Weighted Composite NASA-TLX Ratings per Sub-Condition.

	P101	P102	P103	P105	P106	Mean
1A	13.33	24.00	69.00	31.67	11.67	29.93
1B	3.67	25.33	70.00	41.33	30.67	34.20
1C	8.33	36.00	71.33	56.67	21.33	38.73
2A	79.67	34.67	71.00	50.33	34.00	53.93
2B	83.33	37.00	60.00	49.67	18.33	49.67
3A	41.00	40.67	69.00	34.33	37.00	44.40
3B	35.33	41.33	62.67	49.33	37.67	45.27

Unweighted and weighted NASA-TLX composite scores for each of the three conditions are presented in Figure 7.4, averaged across participants and sub-conditions.

NASA-TLX composite scores range from 0 (very low) to 100 (very high) workload. Condition 1, requiring the participant to monitor one reactor, produced the lowest subjective workload scores (mean weighted NASA-TLX = 34.3). The highest workload ratings were obtained in condition 2, in which the participant was assigned to monitor four reactors (mean weighted NASA-TLX = 51.8). A mix of monitoring and procedure-based control of four reactors produced an intermediate level of subjective workload (mean weighted NASA-TLX = 44.8).

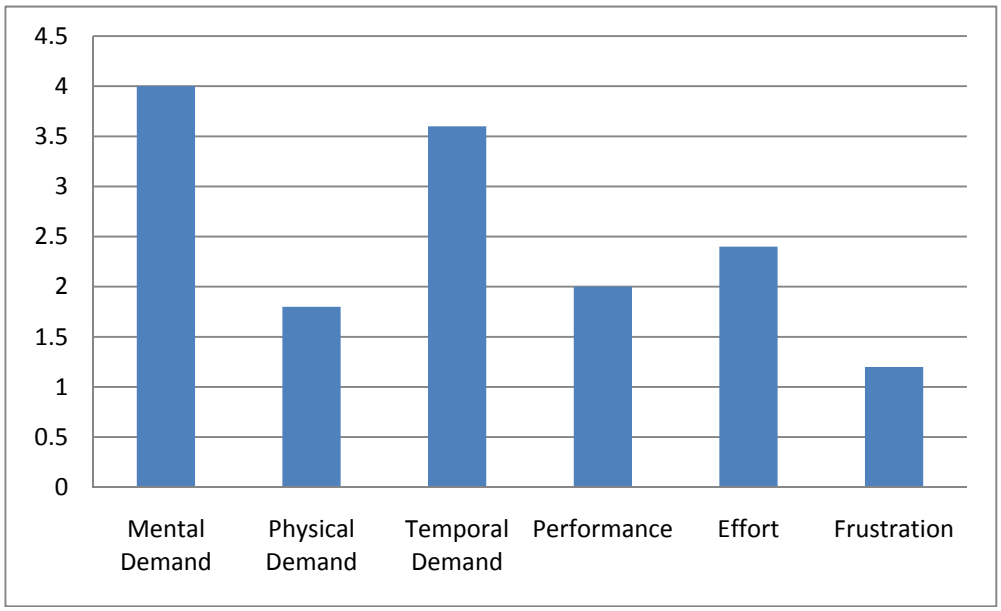

Figure 7.2. Relative Weighting of NASA-TLX Dimensions.

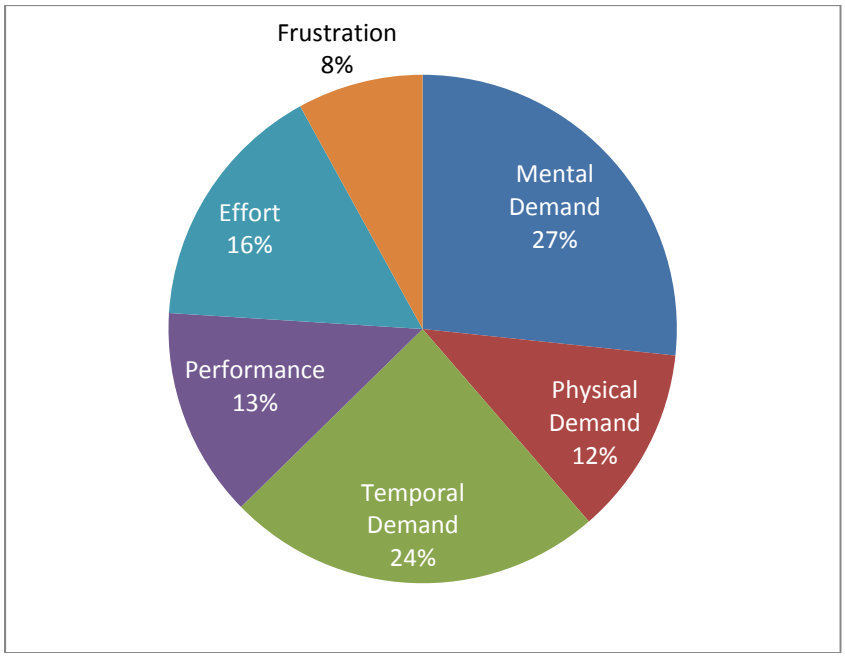

Figure 7.3. NASA-TLX Dimension Weights: Components of Workload.

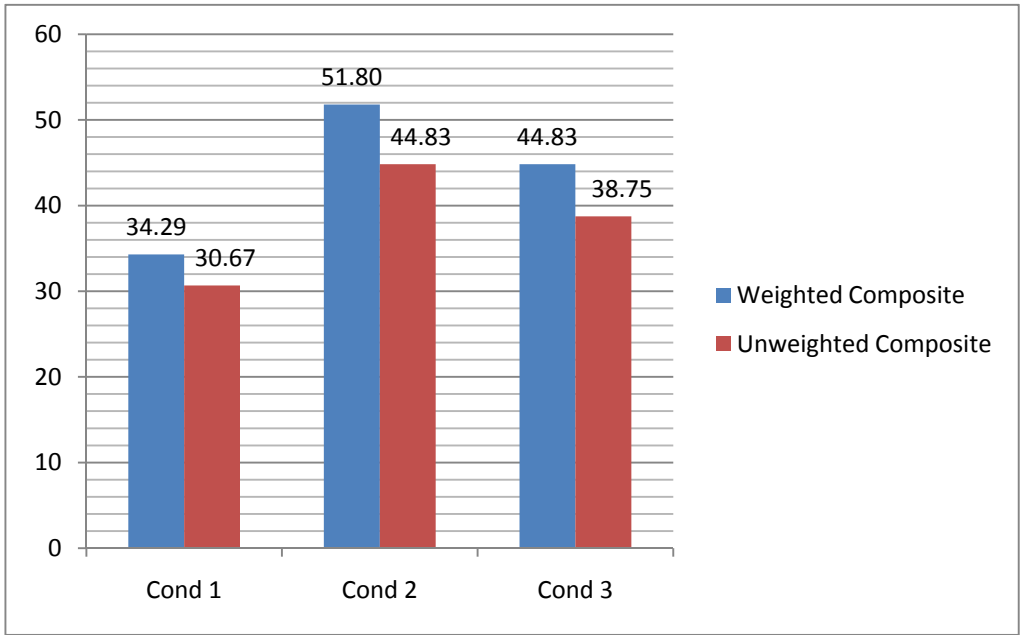

Figure 7.4. Mean Composite NASA-TLX Rating by Condition.

7.2.2 Pupil Diameter

Due to the presence of paper reference materials, logs, and in condition 3, procedures, participants spent a substantial amount of time looking down at the desk surface. During these times, the eye tracker was unable to record eye gaze or pupil diameter. For each sub-condition per participant, we filtered out these periods of invalid data and then averaged the pupil diameter of the right eye for all valid readings. Figure 7.5 presents pupil diameter per sub-condition averaged across all five participants. The largest average pupil diameter occurred during the five-minute at-rest baseline period prior to the three conditions. The mean pupil diameter averaged across all participants per condition was lowest, and very nearly equal, for conditions 2 and 3, which occurred toward the end of the experiment. The order of presentation was balanced for these two conditions. These results did not match expectations, as the well-documented task-evoked pupillary response is the tendency for the pupil to temporarily dilate (i.e., enlarge) with increased cognitive demands. Therefore, with the task demand manipulations in this experiment, and all else being equal, one would expect pupil diameter to be smallest during the baseline recording period. Similarly, considering that participants subjectively rated conditions 2 and 3 to have higher workload than condition 1, one would also expect the demands of conditions 2 and 3 to evoke greater average pupil diameter than the first condition did. Figure 7.6 presents the mean pupil diameter per participant per sub-condition. Here, the values are presented in chronological order, showing a downward trend in pupil diameter over time for most, if not all, participants.

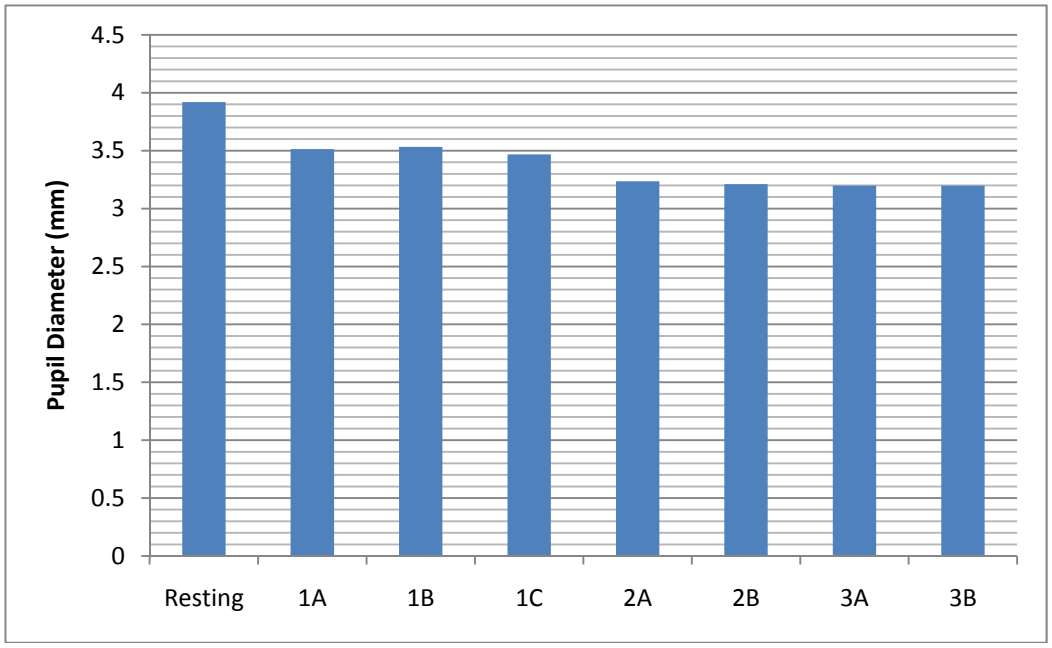

Figure 7.5. Mean Pupil Diameter by Condition, with Order Balanced for Conditions 2 and 3.

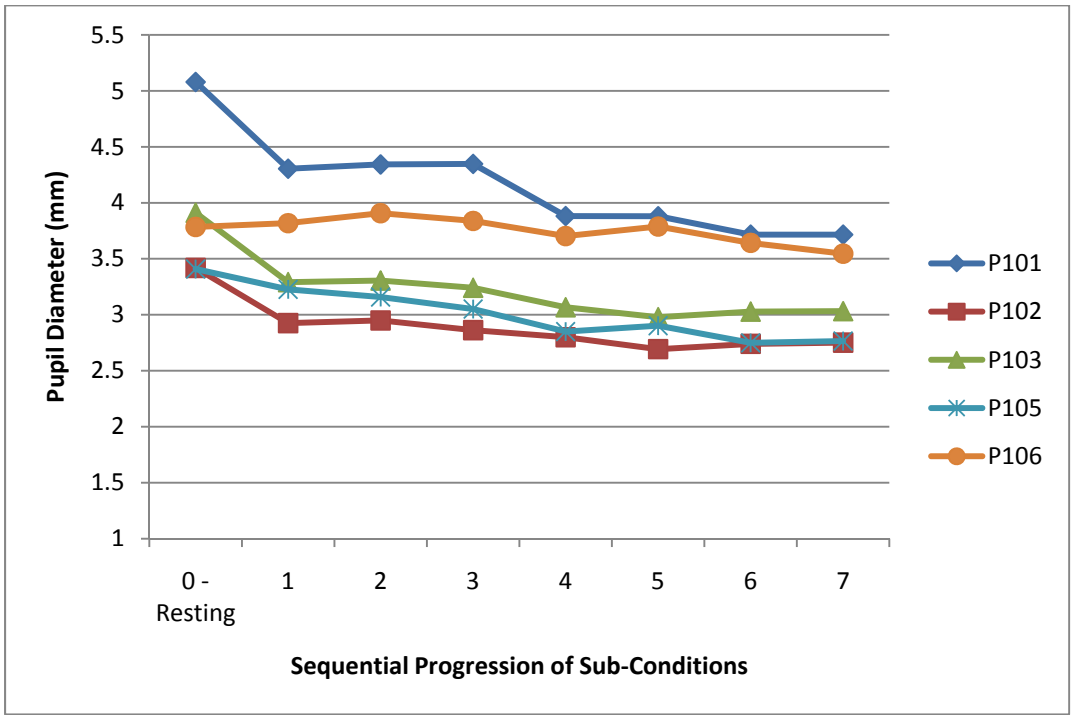

Figure 7.6. Pupil Diameter per Participant over Time, Regardless of Condition Ordering.

Because the pupil diameter results were comparable for conditions 2 and 3, we looked at the pupil data for several five-minute intervals from these conditions with four operating reactors. The specific intervals are discussed in greater detail in section 7.2.3. Averaged across all five participants, the differences between these intervals are small (Figure 7.7). Any effect cannot be confidently reported without greater statistical power, but this pupil diameter analysis follows the trend observed for the heart rate measures (see below). Namely, the physiological effects of the task demands may have been greater during four-reactor monitoring (i.e., "Monitor 4" and "2nd Leak" in condition 2) than during procedure-based control and simultaneous monitoring ("Shutdown Proc" in condition 3).

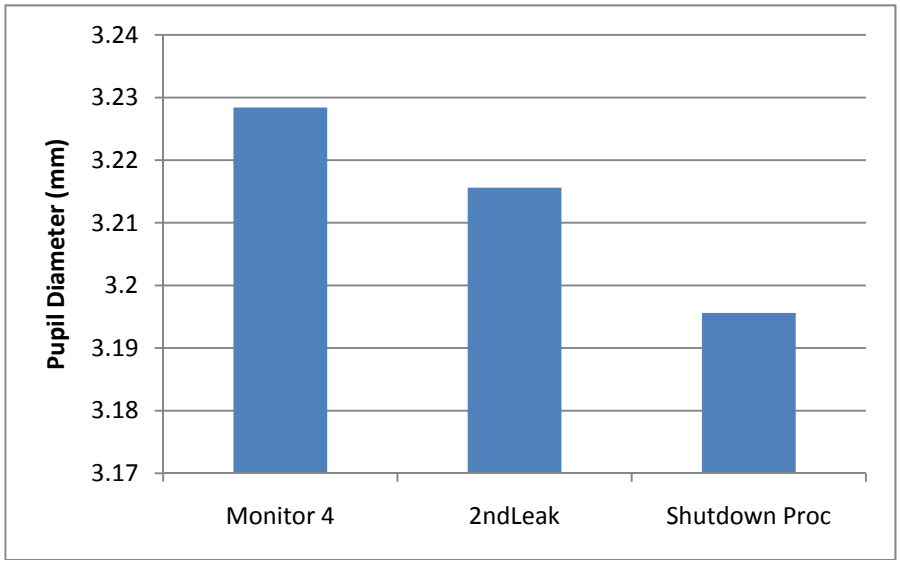

Figure 7.7. Average Pupil Diameter over Five-Minute Intervals.

7.2.3 Heart Rate and Heart Rate Variability

7.2.3.1 Data Analysis

The Polar ProTrainer 5 software provides for convenient analysis of a selected portion of the R-R recorded heart rate data on a number of dimensions. Based on our literature review of heart rate and variability analysis, we looked at the resulting data per sub-condition for a subset of these measures: average heart rate (beats per minute), RMSSD (Root Mean Square of Successive Differences; i.e., beat-to-beat variability), total spectral power, Low Frequency (LF) power (0.04-0.15 Hz) and Low Frequency / High Frequency ratio (LF/HF ratio; HF = 0.15-0.40 Hz). Assuming other factors are fairly controlled across conditions, higher heart rate and lower variability suggest greater workload. A portion of the results are presented below.

We excluded the results of P101 from the aggregate heart rate analyses. This participant reported taking medication for a heart condition, and his heart rate data was markedly different from the other four participants.

Figure 7.8 shows the average heart rate across the remaining four participants per condition. Conditions 2 and 3 had similar results. The resting heart rate during the five-minute baseline condition was noticeably higher, with condition 1 producing the greatest average heart rate. Subjectively, conditions 2 and 3 incurred greater workload than condition 1, as expected. Thus, as with the pupil measure, these unexpected heart rate results seem to reflect effects other than workload. The ordering of conditions 2 and 3 was balanced, but the baseline was always obtained first, followed by condition 1.

Therefore increased fatigue, emotional comfort or task disengagement may have affected heart rate during the course of an experimental session.

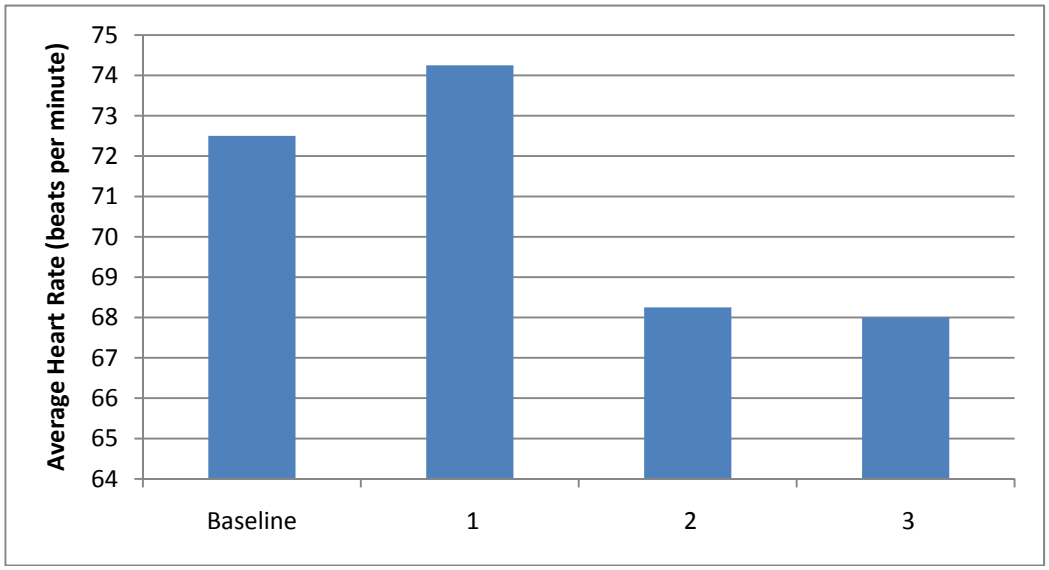

Figure 7.8. Average Heart Rate per Condition, Excluding P101.

The heart rate results per sub-condition for P105 are notable (Figure 7.9). In the post-experiment interview, P105 reported that it was stressful when the reactor power was greater than normal and increasing (e.g., in conditions 1A and 1C). In condition 1C the reactor power was well above the upper threshold. This participant was a former TRIGA reactor operator, and she recounted a time during actual operations when the reactor exceeded the specified power limits without the crew's knowledge, resulting in subsequent problems. Also, during condition 2B a second leak occurred on reactor 2, causing the pressurizer level to drop continually until the end of the scenario.

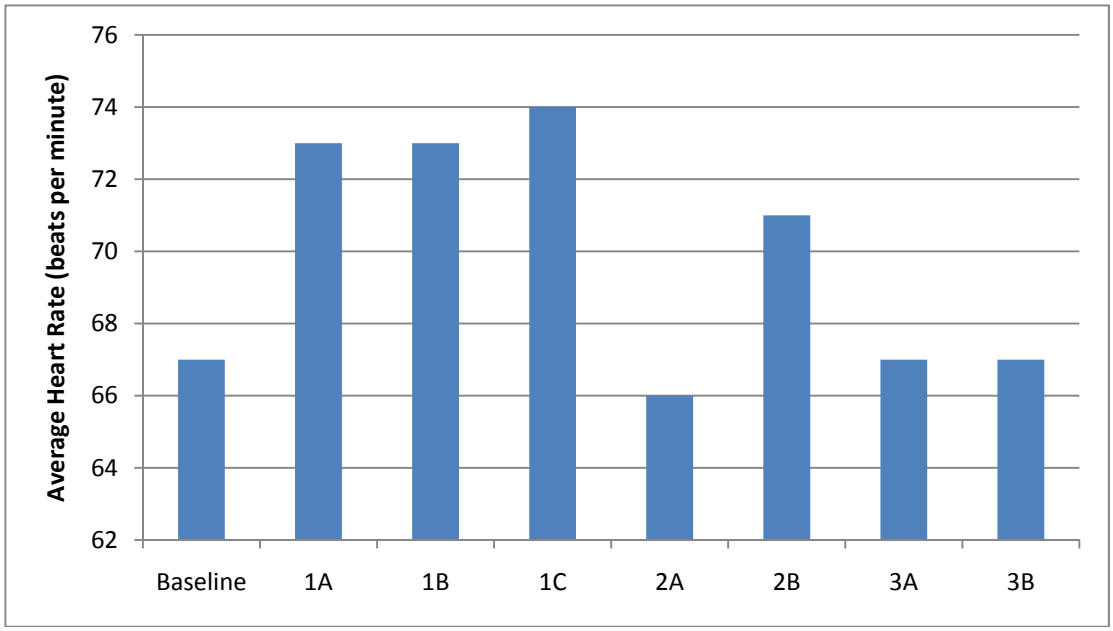

Figure 7.9. Average Heart Rate per Sub-condition for P105.

Figure 7.10 displays the moment-to-moment heart rate recording from participant P105 for condition 1. This graph demonstrates the considerable variability of the beats per minute (e.g., here, from about 60 to about 100 bpm) which may be observed in a seated display monitoring task with minimal physical demands. Although many peaks are inexplicable, some major trends can be linked to simulator events. Such effects on heart rate were more pronounced for this participant than most. The sustained periods of heightened heart rate in sub-conditions 1A and 1C seem to have objectively captured the psychological stress reported afterward by the participant.

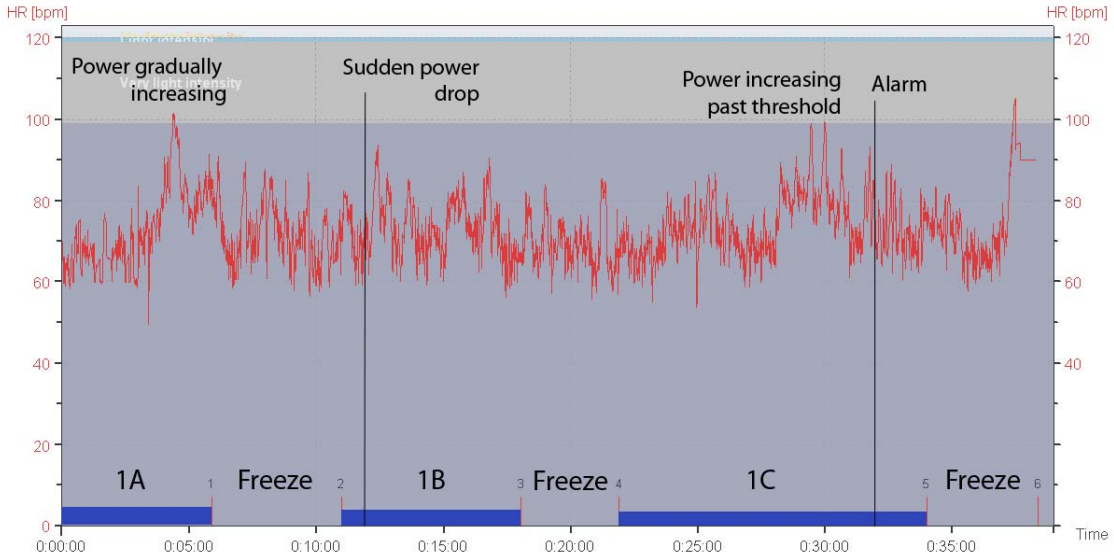

Figure 7.10. Heart Rate for P105 in Condition 1.

We compared the average heart rate per participant between conditions 2 and 3, both of which required the participant to multi-task between four reactors. The within-subjects differences are minimal but follow an interesting pattern (Figure 7.11). In the post-session interviews, participants 101, 103 and 105 rated condition 2 as more difficult, whereas participants 102 and 106 found condition 3 more difficult. In each of these five cases, the individual's heart rate averaged over 25 minutes of simulation was slightly higher for the condition which he or she rated as more difficult. However, these results also match the order of presentation, as each participant rated the condition encountered first as more difficult.

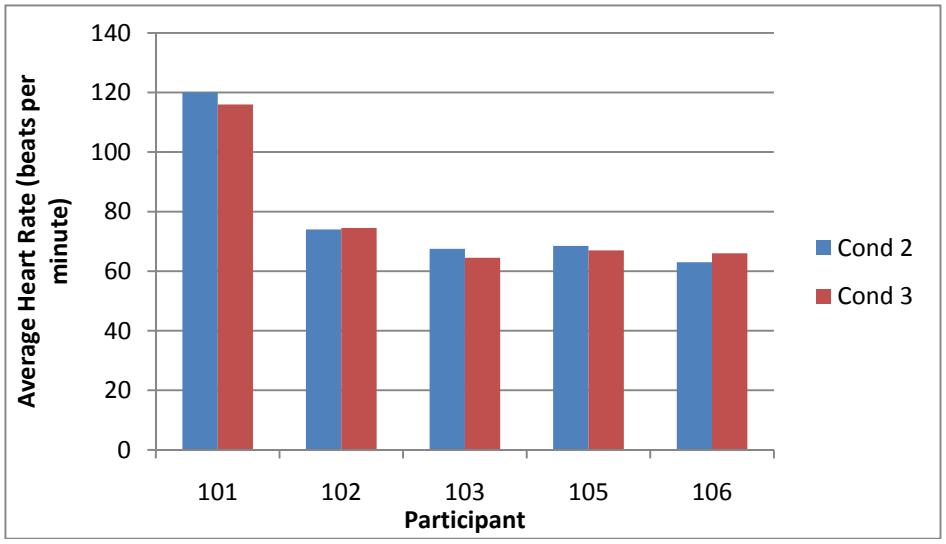

Figure 7.11. Average Heart Rate per Participant for Conditions 2 and 3.

Spectral power is one of many possible measures of heart rate variability, with greater total power reflecting higher variability of the heart rate over time. Low frequency (LF) power (0.04-0.15 Hz) is a portion of this spectrum particularly sensitive to variations in mental effort. Lower power, and thus lower heart rate variability, suggests higher mental workload or effort, assuming other factors are equal. Comparing conditions 2 and 3, LF Power may have followed the same trend as average heart rate to a lesser extent (Figure 7.12). On the other hand, LF/HF ratio was lower in condition 2 than condition 3 for all participants, regardless of presentation order. This may suggest that, generally, greater mental effort was expended in condition 2 (i.e., passive monitoring and logging).

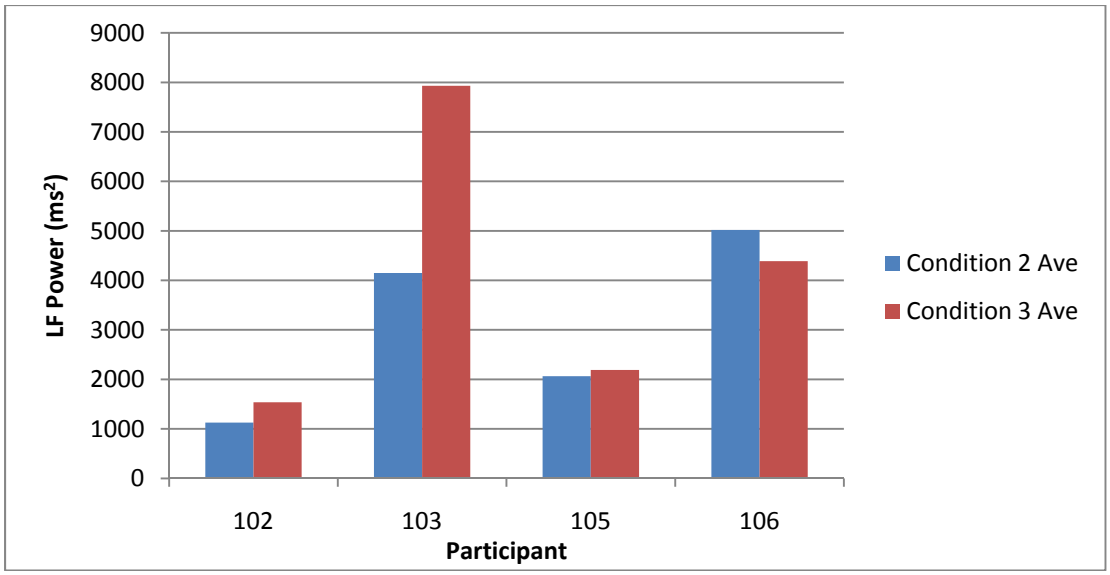

Figure 7.12. LF Power in Conditions 2 and 3, Excluding P101.

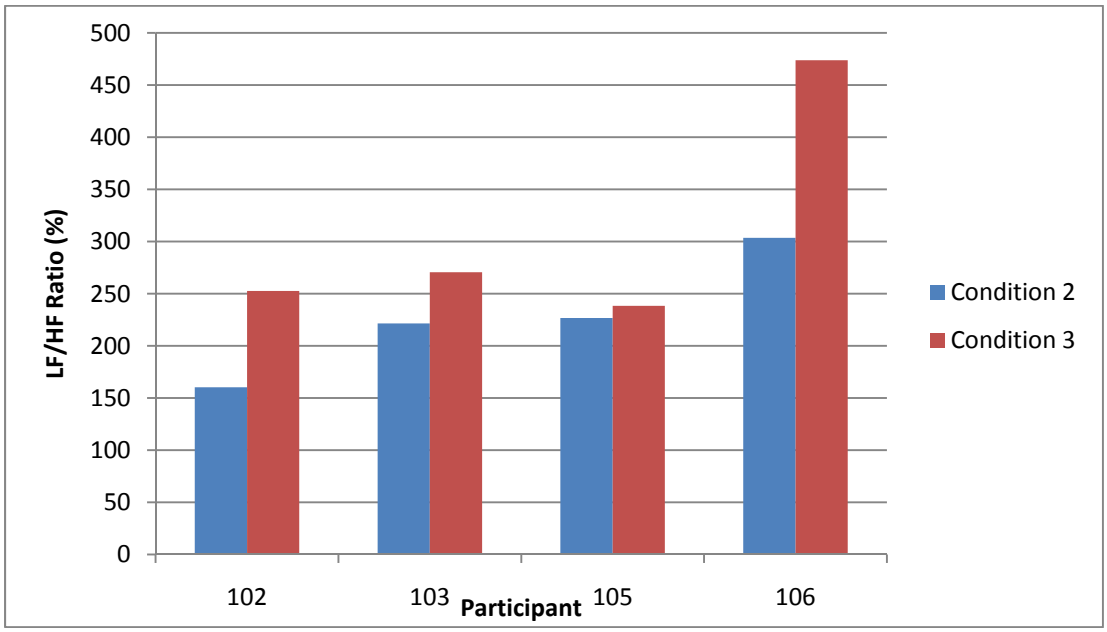

Figure 7.13. Comparing LF/HF Ratio in Conditions 2 and 3. Two participants encountered each condition first.

An advantage, and potential challenge, of the physiological measures is their continuous nature. Whereas a subjective measure of workload such as NASA-TLX is obtained at most every few minutes, there are many samples of heart rate data per second.

This data can theoretically be linked to specific events during the simulated task, which provides additional information, but also requires time-consuming, tedious analysis. Because of the irregularities in the physiological data between the baseline and first condition together, as compared with the second and third conditions together, we chose several representative five-minute intervals from the latter two conditions with four operating reactors for more in-depth analysis of the heart rate data. Five minutes is a common interval for HRV analysis (Healey and Picard, 2005). The three intervals investigated at this level were: Condition 2, simulation time 7-12 minutes (monitoring four reactors, with reactor 4 gradually recovering from an incident); Condition 2, simulation time 18-23 minutes (monitoring four reactors as second leak occurs on reactor 2); Condition 3, simulation time 5-10 minutes (reactor 3 shutdown procedure in progress, reactor 1 load increase in progress). That is, we looked at monitoring of relatively normal operations, monitoring during an emergency, and simultaneous procedure-based control and monitoring, respectively. Again, P101's data was excluded from this analysis.

The maximum heart rate obtained per five-minute interval, averaged across the four participants, varied considerably between the three intervals (Figure 7.14). A higher maximum heart rate seems to indicate greater workload, stress, effort and/or arousal during the onset of the second leak in reactor 2, considering there was little variation in physical activity. This is noteworthy, considering one of the four participants included in this analysis did not report or log the occurrence of the second leak.

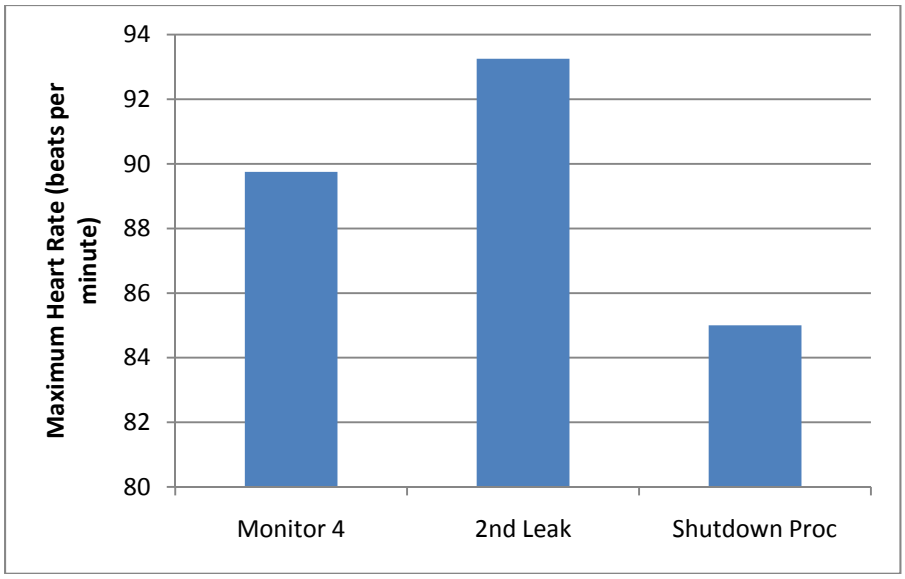

Figure 7.14. Average Maximum Heart Rate During Five-Minute Intervals, Excluding P101.

The average heart rate across the four participants for each five-minute interval also seems to suggest greater workload during the interval containing the second leak, but the effect is less pronounced than for maximum heart rate (Figure 7.15).

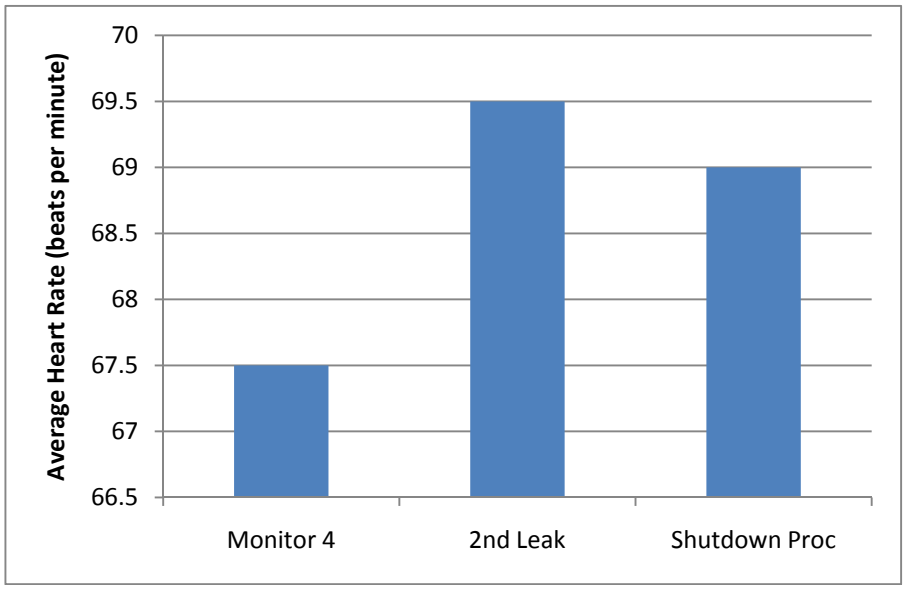

Figure 7.15. Average Heart Rate During Five-Minute Intervals, Excluding P101.

As with the increased maximum and average heart rates, the diminished spectral power during the second leak in condition 2 suggests greater workload, effort or stress during this interval (Figure 7.16). The LF power graph for the three intervals (Figure 7.17) is somewhat different, implying workload was greater during the passive monitoring task than the active control task, but with a minimal difference between monitoring under normal and abnormal conditions.

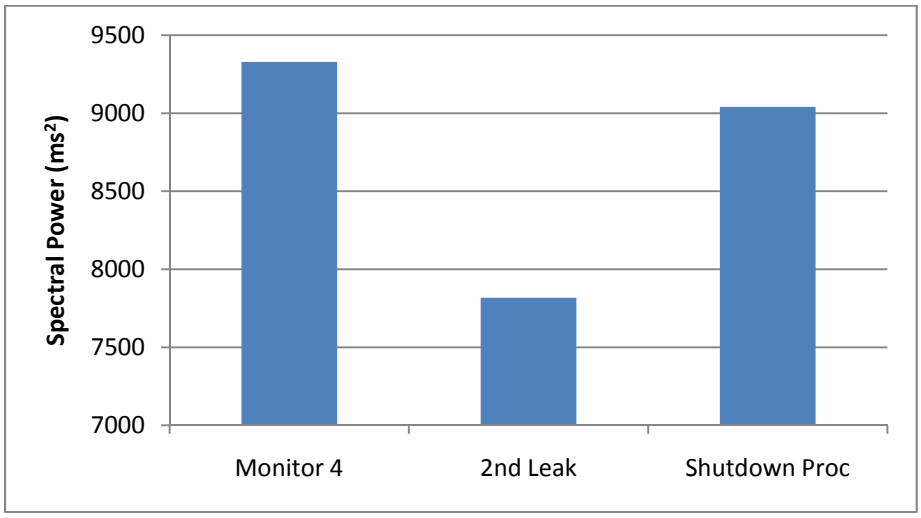

Figure 7.16. Total Spectral Power During Five-Minute Intervals, Excluding P101.

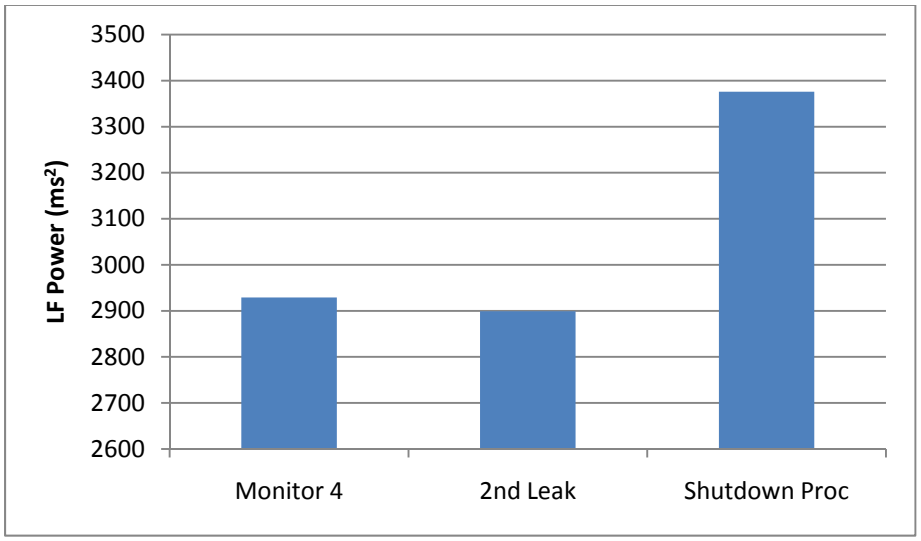

Figure 7.17. LF Power During Five-Minute Intervals, Excluding P101.

Both heart rate and heart rate variability appear sensitive to differences between 25-minute scenarios and shorter intervals within scenarios. In comparing the various measures there are inconsistencies, but some reasonable patterns seem to have emerged. The relationship of the measures for the first condition as compared to the subsequent conditions is troubling and is discussed in Chapter 8. Because of this, we focused primarily on conditions 2 and 3 for the preceding heart rate data analyses.

7.2.3.2 Investigation of Freeze Effects

We also sought to determine whether the simulation freezes for SACRI administration might have an observable effect on the heart rate data. This is an important question, because if the answer were yes, it would suggest possible interference between the measures: it seems reasonable to suspect that the freezes provide an artificial break from the monitoring task demands (see Ha et al., 2007), impacting workload, fatigue, arousal, etc. as compared to actual continuous operations. In visually observing the moment-to-moment heart rate graphs, we did not find obvious, consistent differences between the simulation and freeze periods. In some cases we observed temporarily heightened heart rate around the time of simulator resumption, but this was often within the bounds of the normal variability of the graph during a condition. It is possible the effect would be more pronounced in a higher fidelity simulation, which may induce greater task- and environment-related stress. However, comparing average heart rate between the three sub-conditions and the three freeze intervals of condition 1, we did

observe some interesting trends (Figure 7.18). For participants P101, P105 and P106, all with operating experience, the heart rate tended to drop during a freeze as compared to the adjacent simulation intervals. This suggests the freeze may have provided a momentary break from the monitoring demands of condition 1. The pattern for P102, also with operating experience, is less clear. For P103, who lacked operations experience, the opposite trend is apparent: heart rate heightened during the simulator freezes, indicating the SACRI administration may have been even more demanding than operation for this participant. This is consistent with his freeze times, which were on average the longest compared to the other four, more experienced, participants (average 6.31 minutes vs. 4.44 minutes). Additionally, he indicated that SACRI was a stressful part of the experiment. For conditions 2 and 3, the differences in heart rate between simulator time and freezes appeared to be less pronounced.

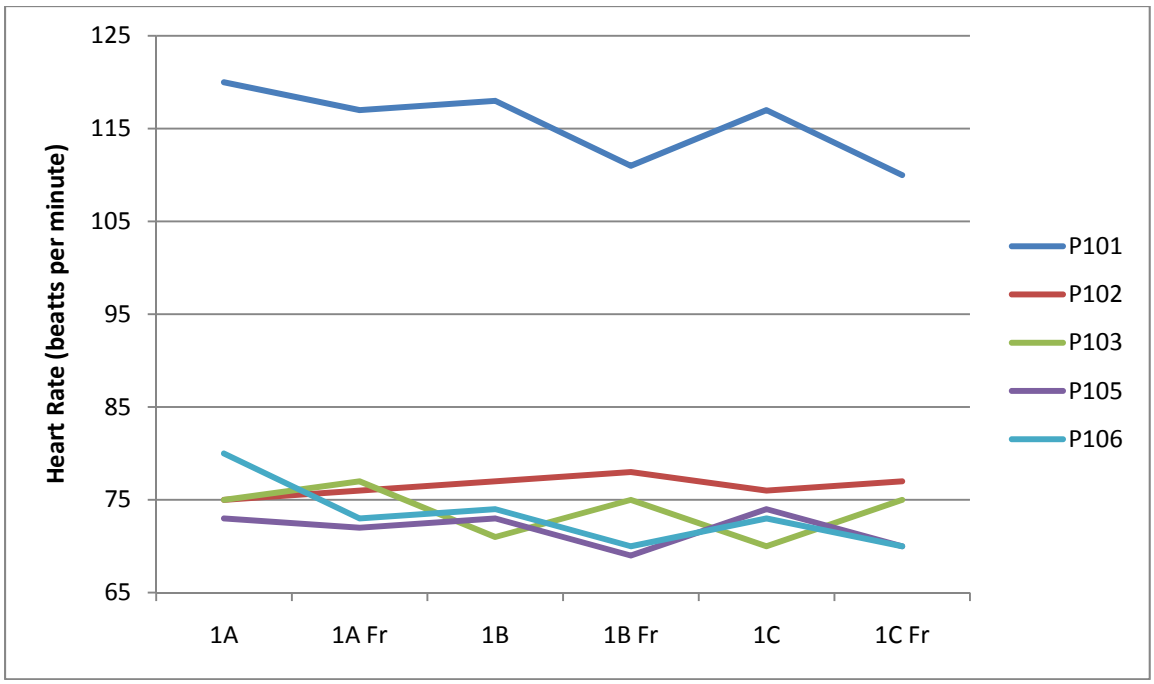

Figure 7.18. Average Heart Rate for Condition 1: Comparing Simulated Monitoring Sub-Conditions and Freezes.

7.2.4 Relating Various Measures of Mental Workload

Although subjective and physiological techniques measure workload at different timescales, it is desirable to investigate the relationships between them, in seeking convergent validity. Most of these analyses were inconclusive for this study, perhaps due in part to the ordering effects on the physiological measures identified above. For example, average heart rate varied little across the wide range of subjective workload reported via NASA-TLX. However, a few noteworthy results are included here.

The pupil diameter data behaved unexpectedly between the resting baseline, condition 1, and the subsequent conditions, likely reflecting other factors besides workload. We therefore compared average pupil diameter per sub-condition with NASA-TLX ratings, but limited this analysis to conditions 2 and 3, for which order was balanced. The results are potentially interesting (Figure 7.19), as a positive correlation between the two dissimilar workload measures may be expected. However, the exclusion of the data from condition 1 resulted in relatively few data points for this scatter plot.

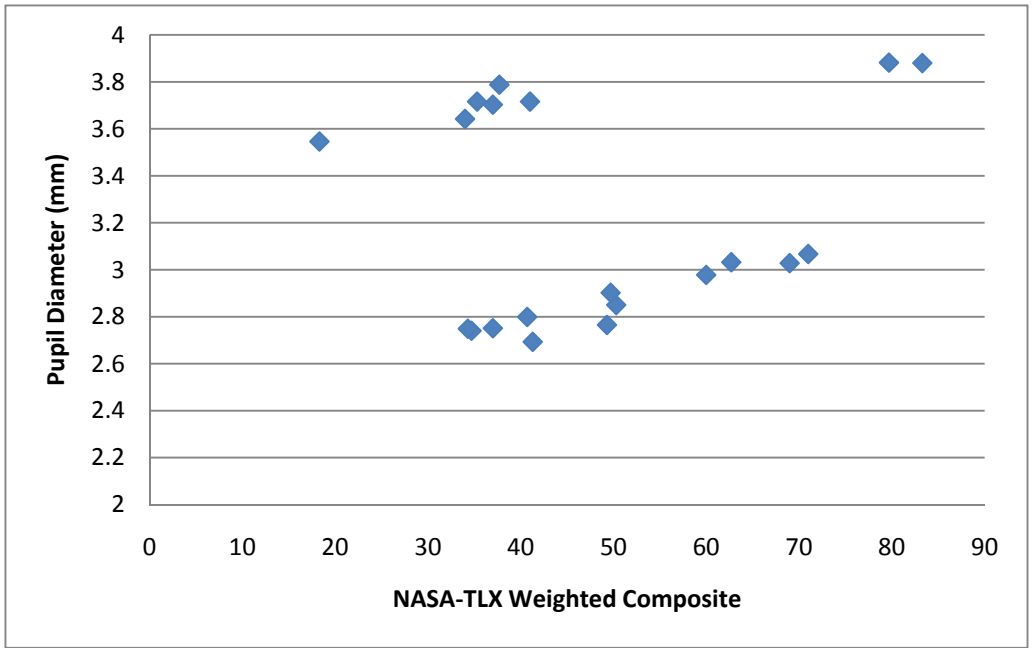

Figure 7.19. Pupil Diameter x NASA-TLX for Conditions 2 and 3.

Figure 7.20 presents six minutes of moment-to-moment physiological data for P105. Specifically, this is heart rate and pupil diameter data from condition 1A. During this segment of condition 1, the reactor power was increasing above 100% of normal operations. As mentioned above, such events were subjectively stressful for this participant. Due to eye tracking issues and the frequent viewing of reference materials on the desk, there are large gaps in the pupil diameter data. However, a sharp peak in pupil diameter is observable midway through the fourth minute, which coincides closely with a clear rise in heartbeats per minute. This is precisely the point at which she reported that core power was at 102.4% and "steadily increasing." Such analyses are tedious, and most results were much less obvious. However, these corresponding graphs provide some evidence for convergent validity of the physiological measures applied: in the very least, heart rate and pupil diameter are demonstrably sensitive to related processes.

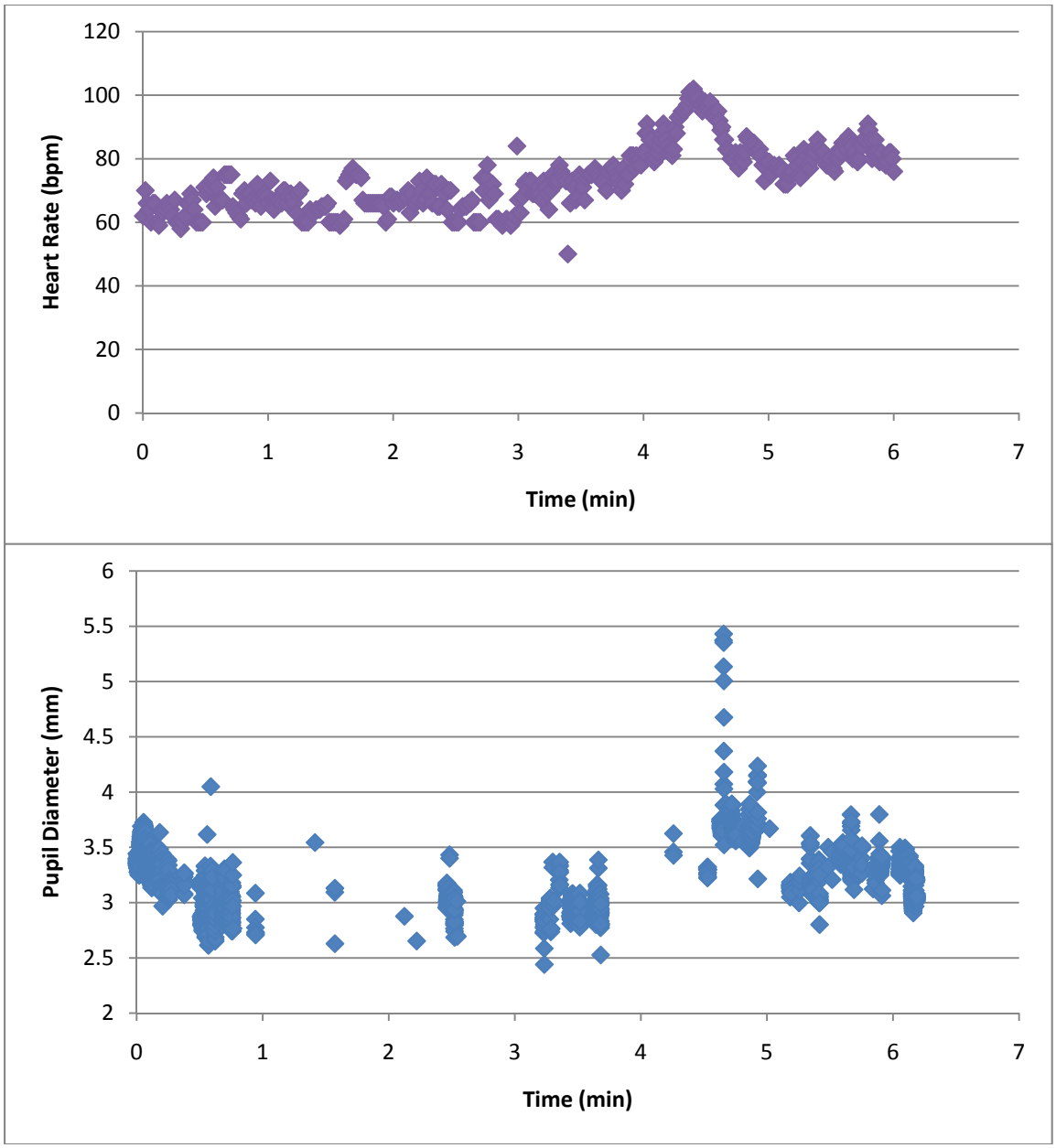

Figure 7.20. Relating Physiological Measures of Workload for P105, Condition 1A. Top: Momentary Heart Rate (beats per minute). Bottom: Pupil diameter (mm).

7.3 Situation Awareness

7.3.1 Modified SACRI

As the freeze times and questions selected were not fully randomized, the results of SACRI should be interpreted with caution. Averaged across freezes and participants, response accuracy to the SACRI questions was highest in condition 2 and lowest in condition 3 (see Figure 7.21). In this graph, a score of 1.0 means that all five participants answered all SACRI questions correctly for a given condition. For individual applications, accuracy ranged from 7.5 to 12 correct responses out of 12 questions, but scores for 26 out of the 35 questionnaires fell between 9 and 11. This either means situation awareness was remarkably consistent between conditions, or the instrument is not very sensitive with this type of analysis.

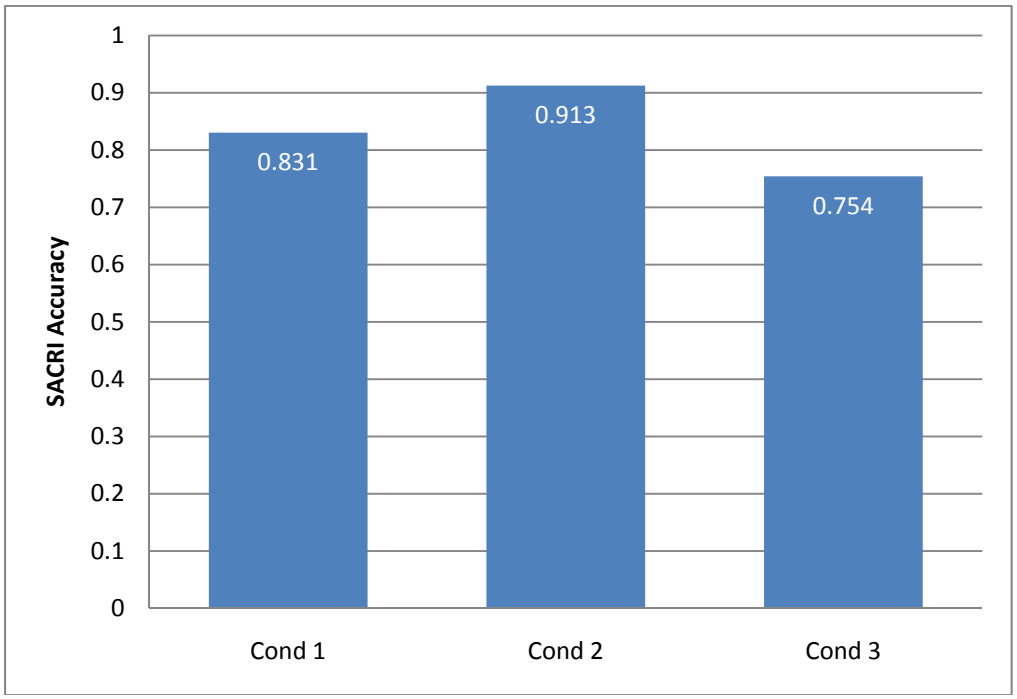

Figure 7.21. Objective Situation Awareness: Overall SACRI Accuracy per Condition.

7.3.2 Subjective Situation Awareness Rating

For each sub-condition, the participant rated his or her own situation awareness between 1 (Emergency) and 5 (Excellent). This scale had poor sensitivity, as most ratings were either 3 (Fair) or 4 (Good). Out of the 35 subjective ratings, only one was 1 (Poor), and three were 5 (Excellent). Participant 106 marked "Good" situation awareness for all seven sub-conditions. The mean subjective situation awareness varied little between conditions, as shown in Figure 7.22, perhaps due to the fact that this simple scale lacked sufficient range for precise ratings. This should not be surprising, because the scale was intended to identify and explain exceptional circumstances (i.e., outliers), not subtle differences in subjective awareness. However, it is surprising that the mean rating was slightly higher in the third condition (procedure-based control with four modules) than the first condition (monitoring one reactor module). The mean normalized subjective ratings per sub-condition are compared with mean SACRI accuracy in Figure 7.23.

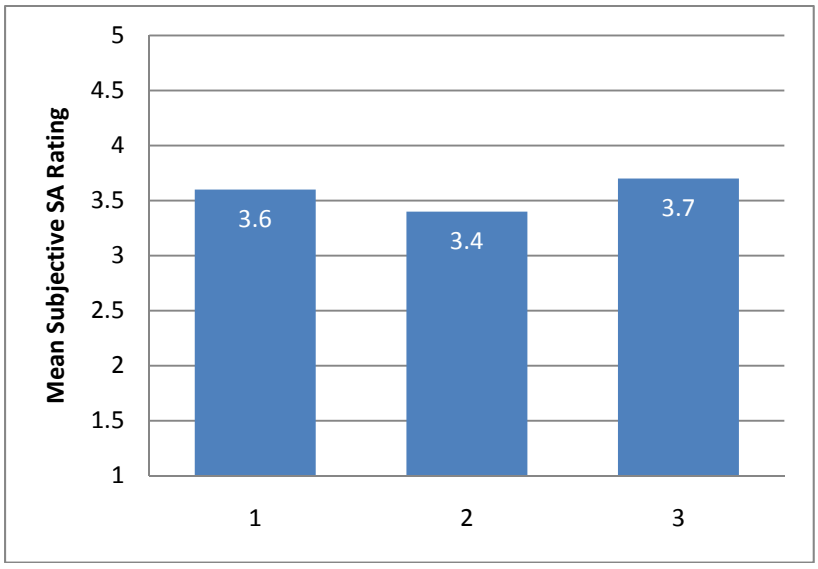

Figure 7.22. Mean Subjective Situation Awareness per Condition.

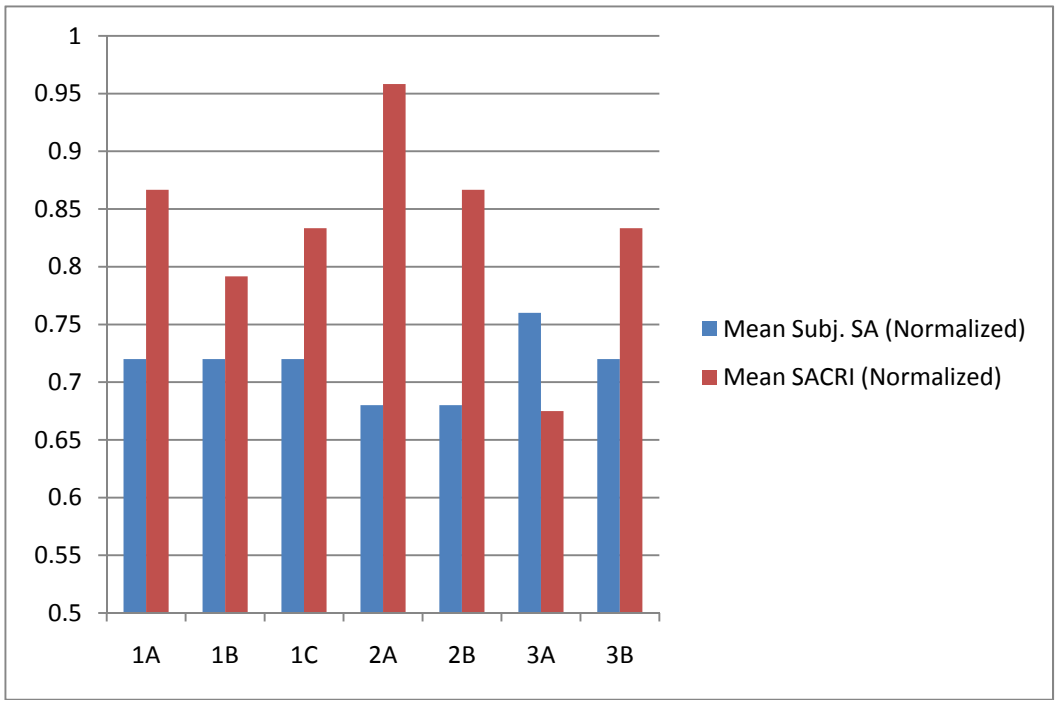

Figure 7.23. Comparing Normalized Subjective and Objective Situation Awareness by Sub-Condition.

7.3.3 Eye Tracking

We performed two primary types of high-level analysis of the eye tracking data for investigating situation awareness. First, composite heat maps were generated via the Tobii Studio software for each condition, combining the data from all five participants. These graphical representations may indicate where participants collectively directed most of their visual attention, as well as what information they may have neglected. Second, we also inspected various time-sequenced gaze plots in the Tobii Studio application for the sake of learning about operator scan strategies. These analyses are therefore qualitative in nature, but nevertheless interesting and enlightening.

We did not attempt to analyze the eye tracking data at the user interface component level. Because of the geometry and size of the display area tested, such

analysis would likely meet with hurdles due to parallax error and individual calibration inaccuracies. However, even in cases when distortion of the gaze data was observed, we could still determine which of the eight flatscreen displays was being attended. It should be reiterated that there were two flatscreen displays per reactor, stacked vertically, with the main mimic display on the bottom and trend plots on the top. Reactors 1 to 4 were arranged horizontally, from left to right.

With a heat map visualization, red represents the most frequently attended display areas. Figure 7.24 shows the heat map for condition 1, in which the participant monitored reactor 3 only. Figure 7.25 is the heat map for condition 2, in which all four reactors were operating, and reactors 2 and 4 had issues. The participant also carried out a brief corrective action control for reactor 4 in this condition. Figure 7.26 shows the gaze results for condition 3. In this condition, reactors 1 and 3 were controlled via procedures. Reactor 4 was also to be controlled, but emerging issues caused a hold in the procedure for the rest of the scenario. The operator was also instructed to monitor reactor 2 and fill out a periodic log for this reactor only; it was in steady-state and had no issues. At this level of analysis, it is clear that overall, participants attended to all operating modules, but more heavily to those with issues or assigned procedures. Also, it is clear that participants relied much more heavily on the lower screens (i.e., the “mimics” with parameter values) than on the trend plots on the upper screens.

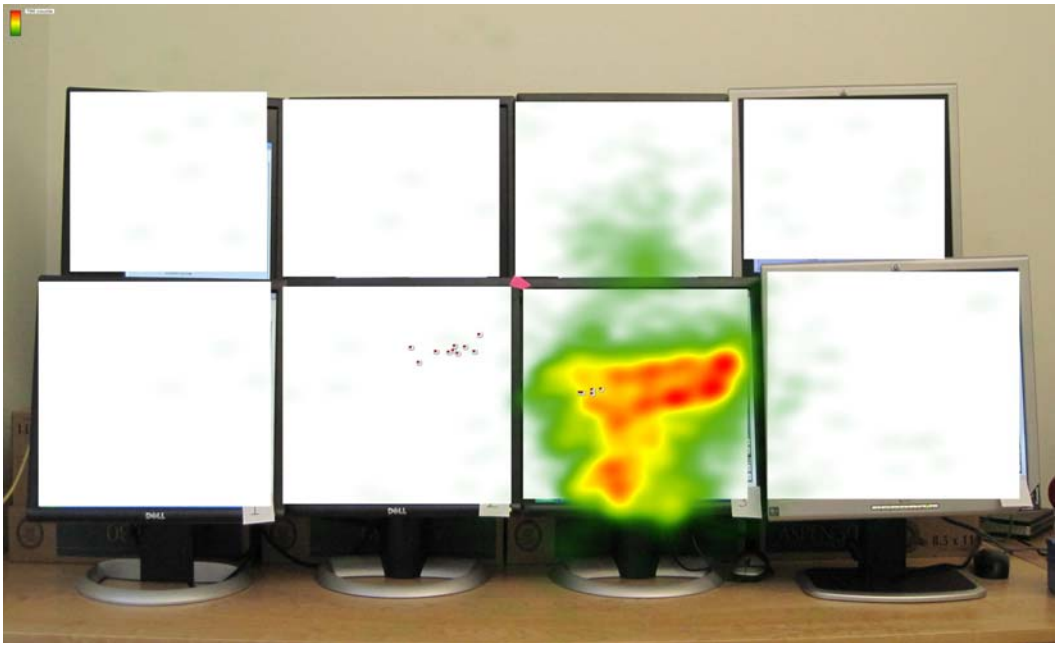

Figure 7.24. Composite Heat Map for Condition 1. Only the screens for reactor 3 were powered on.

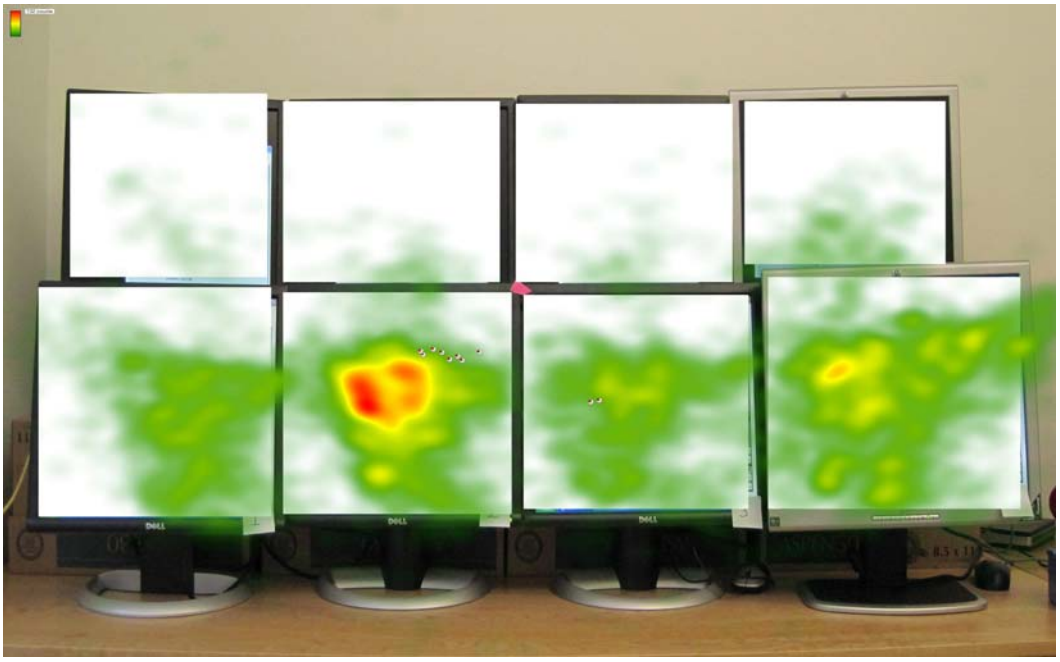

Figure 7.25. Composite Heat Map for Condition 2.

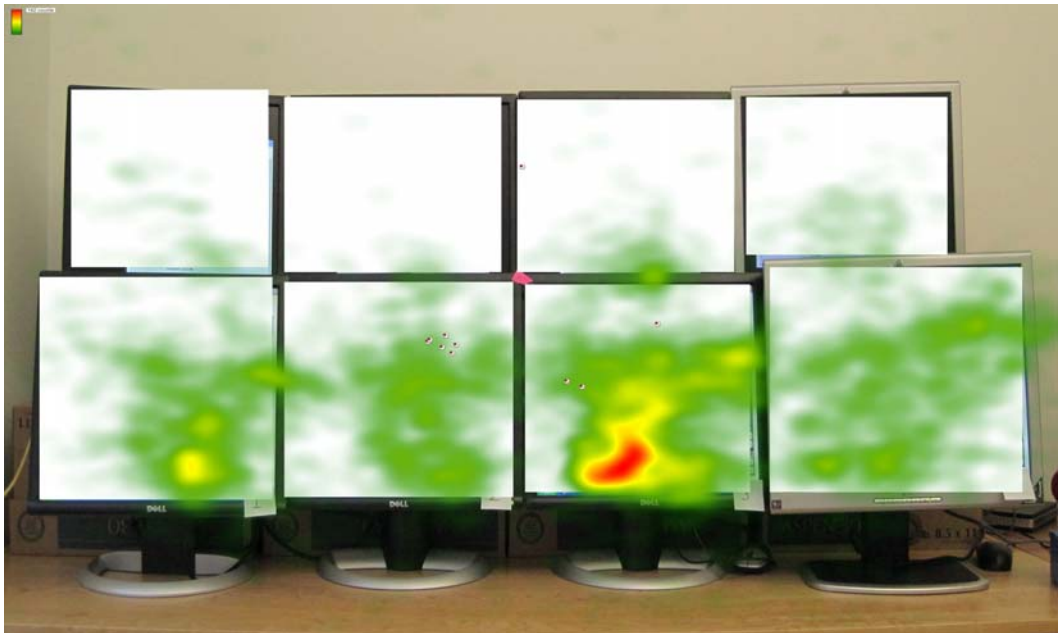

Figure 7.26. Composite Heat Map for Condition 3.

A gaze plot presents the sequence of visual fixations as a set of numbered circles. The size of a circle encodes the gaze duration at that location, and consecutive fixations are connected by a line segment. The gaze plots provided additional information at the individual, sub-condition level. For example, comparing Figures 7.27 and 7.28, it is clear that participant 101, an experienced Navy reactor operator, used the trend plot information consistently in condition 1A, whereas participant 103, with no operations experience, did not use this screen at all.

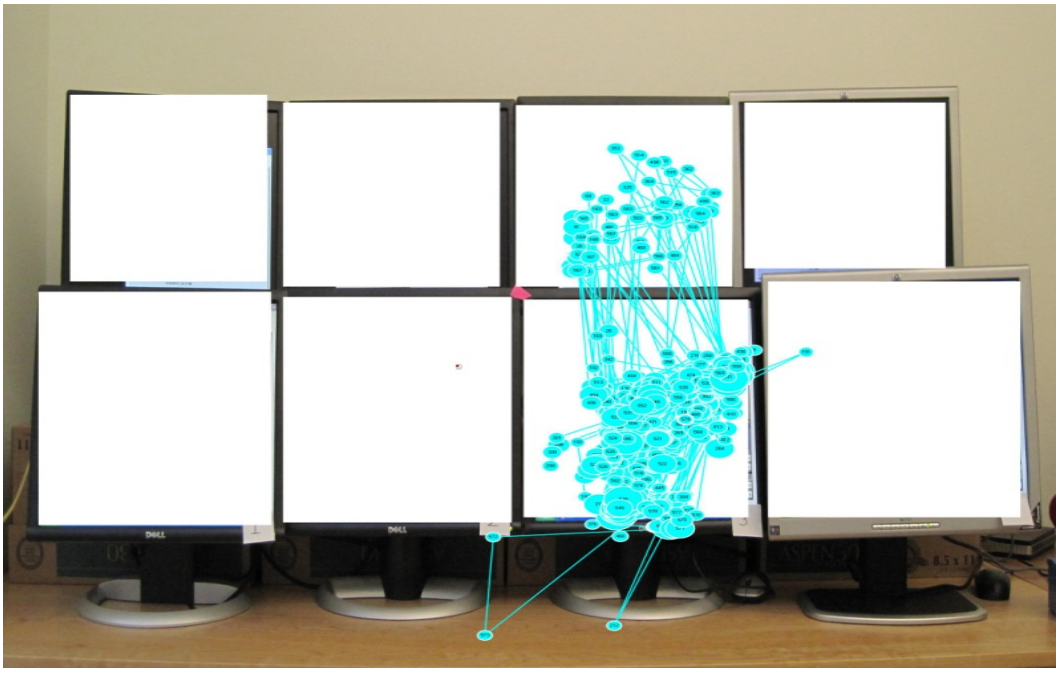

Figure 7.27. P101 Gaze Plot for Condition 1A.

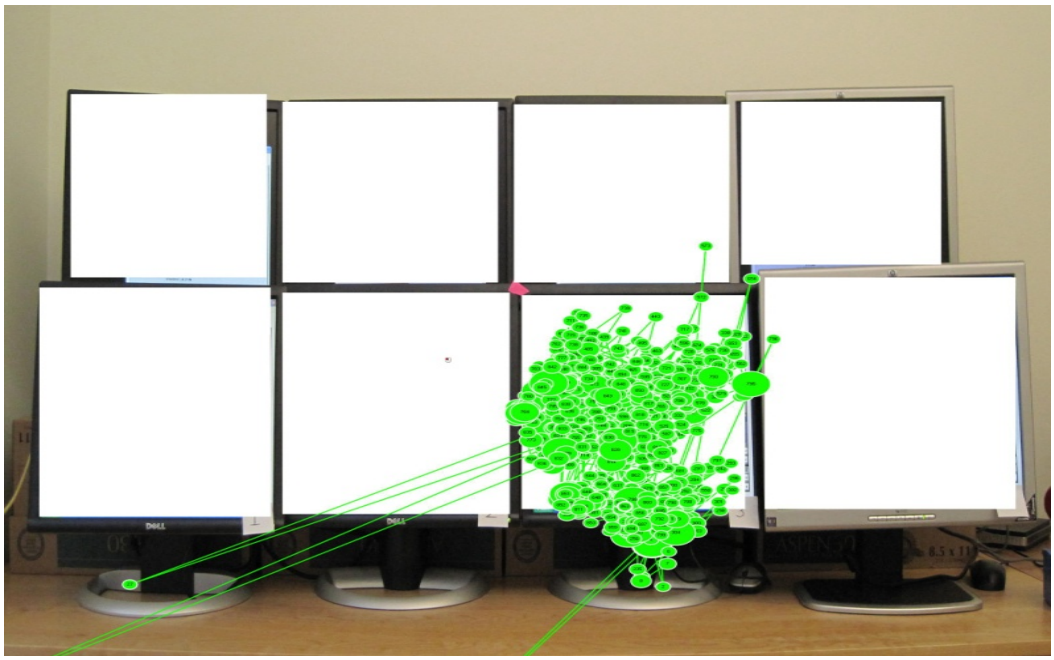

Figure 7.28. P103 Gaze Plot for Condition 1A.

Although we recommended a scan pattern for covering the parameters of interest within a single reactor mimic, and specified a left-to-right higher level scan pattern from reactor to reactor in the operating rules, some participants reported that they found it more expedient to check a single parameter at a time across all four reactors. This strategy is visible in Figure 7.29, where there are many horizontal connections between corresponding locations in adjacent reactor mimics. This image also exemplifies the distortion which we found in some recordings: clumps of fixations suggest separate reactors, but they do not map well to the reference background. Figure 7.30 demonstrates a similar strategy of glancing briefly at each trend plot in succession, for "status at a glance," with attention prioritized to reactor 2.

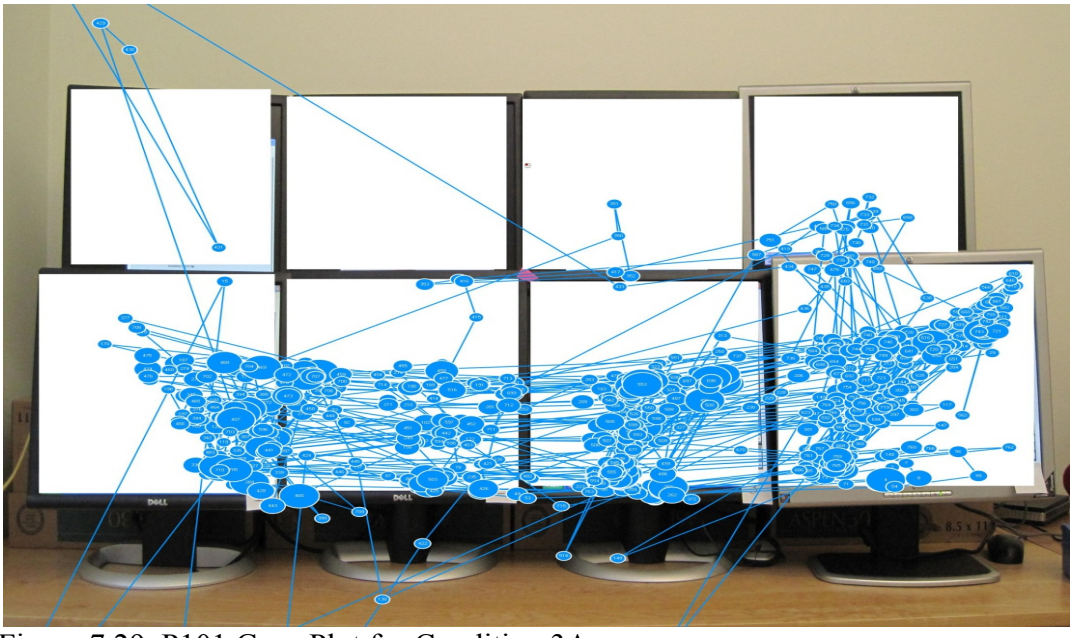

Figure 7.29. P101 Gaze Plot for Condition 3A.

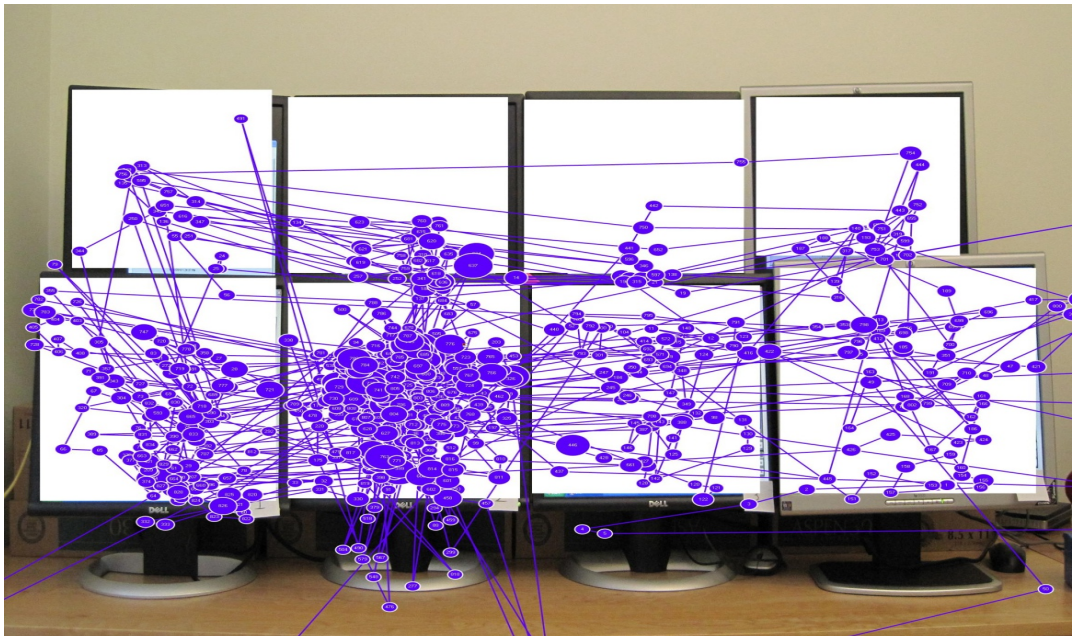

Figure 7.30. P102 Gaze Plot for Condition 2B.

We performed an analysis of the eye gaze data per participant for the same five-minute intervals from conditions 2 and 3 that we investigated for the heart rate and pupil diameter measures. A five minute interval is also desirable for eye gaze analysis in this study, because the operating rules specified that the operator limit continuous monitoring of a single reactor to one minute at a time and to monitor all four reactors within any given four-minute window. Therefore, for ideal monitoring performance, a portion of the participant's visual attention should be clearly devoted to each reactor within any five minutes analyzed.

For condition 2, which consisted of passive monitoring of the four reactors and logging tasks, the gaze plots suggest that all participants did a reasonable job of distributing their visual attention across the four reactors. There was a tendency to allocate more attention to reactors under abnormal conditions, but even during such

events, participants were disciplined to monitor other reactors behaving normally (e.g., Figure 7.31, showing P102's gaze data for five minutes starting just before the second leak began on reactor 2).

Figure 7.31. P102 Gaze Plot for Condition 2, 18-23 minutes (Second leak).

However, this was not the case in condition 3, in which participants were assigned control procedures for several reactors. Some participants tended to focus on the reactors under procedure-based control to the point of neglecting other reactors. This was observed in the eye gaze data for a Navy reactor operator (P106) as well as less experienced participants (e.g., P103; Figure 7.32). We also observed this tendency in the final steps of the reactor 3 shutdown procedure, a five-minute interval we did not consider for the heart rate analysis.

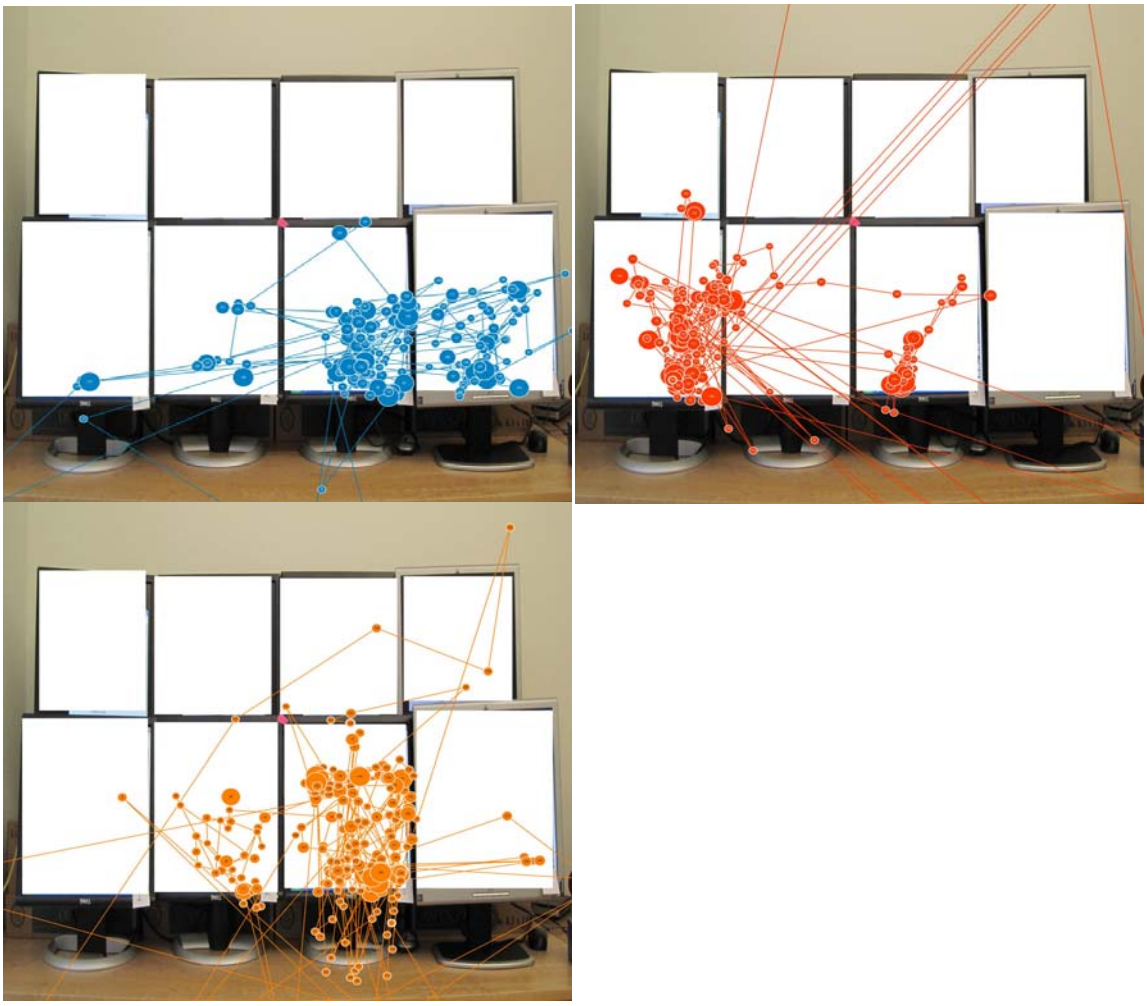

Figure 7.32. Five Minute Intervals of Gaze Data from Condition 3 Exhibiting Problematic Monitoring. Top left: P106 neglects reactor 1 (5^{th}-10^{th} minute). Top right: P103 neglects reactors 2 and 4 (5^{th} – 10^{th} minute). Bottom: P102 neglects reactors 1 and 4 while completing the shutdown procedure for reactor 3 (25^{th} – 30^{th} minute).

Additional images are included in the discussion of the eye tracking results below, as an example of how analysis of such data can be misleading or inconclusive.

7.4 Insights from Post-Session Interviews

Following data collection, we asked participants about tasks demands, their experience with the physiological measures and survey instruments, the simulator

interface, and demographic and background information. Responses are included throughout the proceeding discussion of the results as applicable. Other noteworthy responses are briefly summarized here.

All five reported using the trend plots, but on average, participants estimated they only spent 7.5% of task time looking at the plots. This suggests they were able to obtain "status at a glance" from this type of information visualization.

We were particularly interested in the task demands of monitoring four reactors. Participants found this task somewhat demanding on several dimensions, but we also found substantial individual differences. P105 reported that being assigned four reactors under the simulated conditions was sometimes "overwhelming," to the point that there were "too many things going on, so I had to care less about each [reactor]." On the other hand, P102 reported that he varied his scan pattern to prevent boredom during monitoring. P102 and P106 estimated that the difficulty of monitoring four reactors was about twice that with one reactor; we did not ask all participants about this.

Table 7.3. Task Demands for Monitoring Four Reactors.

Question	**Min. Response**	**Ave. Response**	**Max. Response**
In terms of keeping your eyes focused, how difficult was it to monitor four modules? (1- very easy to 10 - very hard)	3	5.3	8
In terms of keeping your mind/attention focused, how difficult was it to monitor four modules? (1- very easy to 10 - very hard)	3	6.1	8.5
In terms of "keeping everything straight" (i.e., not confusing modules), how difficult was it to monitor four modules? (1- very easy to 10 - very hard)	2.5	6.1	9

8. DISCUSSION

8.1 Task Performance

From Figure 7.1, it is clear that Navy reactor operators (RO's) were disciplined and prompt in maintaining the logs for all three conditions. We observed that the Navy RO's set up the log entries ahead of time as reminders to themselves. The others relied on memory. It should be noted that Navy RO promptness on logging may be overestimated, because these operators filled out the timestamps ahead of time using the precise time the log entry was due. However, these operators were also superior in terms of log completeness.

Participants with a non-Navy background clearly neglected the periodic log in condition 3 (i.e., during procedure-based multi-tasking). These participants may have become engrossed in the procedures for reactors 1, 3 and 4. It is also possible that they became complacent with reactor 2, which had no issues during this scenario. Alternatively, they may have exhibited task shedding (of logging) in order to manage the workload under complex multi-tasking. The fact that participants generally rated mental workload lower during condition 3 than condition 2 suggests that the neglect of the log in condition 3 may have been due to attentional tunneling or lack of awareness of time rather than perceived overload. Without prompting, P102 noted that he tended to focus on those reactors with issues. It is also possible that the less stringent time requirements for the logging in condition 2 contributed to the apparent differences in logging performance between conditions 2 and 3.

For most participants there was not a significant drop-off in log performance from monitoring one to four reactors, even with four times the paperwork, which implies the participants were not overloaded when monitoring four reactors. This is reinforced by the observation that the SACRI measure of situation awareness suggested good awareness during four-reactor monitoring, as well. However, there is a serious performance decrement in logging when the participants were carrying out procedures (i.e., condition 3). This shows that it is important to measure performance, as participants may report lower subjective workload (e.g., NASA-TLX in condition 3 vs. 2), but they may not be completing the task as assigned in order to "lighten the load." It is important to measure the performance of the various tasks in multi-tasking conditions, but it is hard to combine this into a meaningful composite score because of varying priorities between individuals and over time (Veltman & Gaillard, 1996).

8.2 Mental Workload

8.2.1 NASA-TLX

Overall, the composite NASA-TLX scores for the monitoring tasks (conditions 1 and 2) seem somewhat high. Condition 3 was designed to require diverse multi-tasking and thus to incur relatively high workload. In the long-term monitoring of highly-automated systems, underload (i.e., undesirably low workload) is a potential concern, but that was certainly not the case in this experiment. The results are reasonable, though, because the participants only accumulated 75 minutes of simulator time, instructions and training were minimal, and the operating concept and display interface were relatively unfamiliar to the participants. Also the required logging frequency and the incident rate

were artificially high. It is interesting, though, that relatively high subjective workload was found in the monitoring conditions despite the fact that for the most part the participant had no control over the system.

On average, mental demand and temporal demand combined accounted for half of the weighting in the composite NASA-TLX scores. In a long-term monitoring task, temporal demand is typically expected to be very low, but it may have been artificially high in this study due to the five-minute periodic logging requirement and the high rate of abnormal incidents. However, participants also mentioned that the need to divide attention between multiple reactors and to track many parameters contributed to temporal demand. In the literature, such sedentary supervisory control tasks are typically described as having minimal physical demand; in interviews, participants attributed the physical demand in this experiment to the logging tasks. The instructions phase of the experiment already took a significant amount of time, but future studies should explain the meaning of each NASA-TLX dimension more carefully. Perhaps this would have resulted in a lower physical demand weighting.

For each condition, the actual simulator time only lasted 25 minutes and there were few periods of completely normal system operation with operator inactivity. It seems likely that frustration would be weighted more heavily in long-shift monitoring tasks.

It is also interesting that the composite NASA-TLX scores were higher in condition 2 than condition 3. In condition 2, the task consisted of monitoring and logging. Condition 3 also included procedure-based control of multiple reactors and was intended

to be more taxing. In the interview, three of the five participants ranked condition 2 as more difficult than condition 3; the other two participants reversed these. One participant noted that there were more issues in condition 2, but otherwise condition 3 would have been more difficult. Participant P101 attributed the relatively heavy workload in condition 2 to the frequency of the periodic logs: the operator was asked to manually record seven parameter values per reactor every five minutes. This participant felt that fifteen minutes was more appropriate for such periodic logging. P102 felt that the five-minute log period was unnecessarily demanding and inhibited his ability to monitor the system. It is possible that workload ratings were lower for condition 3 because of inadvertent task shedding; we found some evidence of this in the logging performance and the eye gaze data for condition 3.

The NASA-TLX numerical ratings probably cannot be reliably generalized to real-world systems, as the interface design, operator expertise and logging philosophies would be quite different. Another challenge in our simulated setup was to replicate emergency conditions to the point of inducing psychological stress. This workload factor would have a greater impact in actual operations, and possibly even in a more realistic simulated control room with highly trained operators.

8.2.2 Pupil Diameter

Whereas mean pupil diameter per sub-condition was expected to correlate positively with task difficulty, this measure seems to have been more sensitive to other effects. Participants rated the conditions with four reactors as more difficult than the

condition with only one, but the pupil data does not reflect this. The at-rest baseline was always presented first, and then was followed by the first condition, because the latter was intended to give the participants additional time to become familiar with the reactor simulator prior to performing tasks with multiple reactors. The presentation order of conditions 2 and 3, both featuring four operating instances of the reactor simulator, was balanced across participants, but with only five participants total and an uneven number of participants, there are insufficient grounds to statistically compare pupil diameter as a workload measure between these conditions.

The nature of the pupil diameter data was unexpected, but there are several reasonable explanations. It is perhaps possible that the frequent gaze changes between the displays and the papers on the desk surface had some effect on pupil diameter or the eye tracker estimation, compared to the at-rest condition; we lack a deep enough understanding of the anatomy of the eye to support or refute this hypothesis. It is more likely that the experimental session lasted long enough that time had an effect on mean pupil diameter. It fits with existing research that pupil diameter either decreased as participant fatigue increased (see, e.g., Morad et al., 2000) or emotional arousal decreased (see, e.g., Bradley et al., 2008). It is likely that subjective stress or emotional arousal decreased during the session as the participant become more comfortable with the test environment or more bored with the assigned tasks. In post-session interviews, some participants reported being “more tired” at the end of the experiment (P102, P105), and some also reported being “more bored” as compared to “more engaged ” (P102, P106), despite the increased difficulty compared to condition 1.

Another likely explanation, in retrospect, is that we failed to fully control luminosity levels between conditions: the smaller average pupil diameter in the later conditions may primarily reflect the fact that more LCDs were turned on (i.e., all eight displays were in use in conditions 2 and 3, as opposed to zero in the resting baseline condition and two in the first condition). The lab setup included closing the blinds to control for variations in exterior ambient illumination, but we did not consider the effect of turning various displays on and off during the study. In conditions for which some displays were not used, these should have remained on with a gray background comparable to the average luminosity of the user interface screens. However, the fact that the heart rate measures, which are fully independent of screen illumination, exhibited a similar pattern between condition 1 and the remainder of the experiment, suggests that screen illumination does not entirely account for the differences.

Because of this shortcoming in the experimental design, the only two conditions for which pupil diameter can be meaningfully compared are 2 and 3, as both used all eight display screens, with minimal luminance variation between displays. Across all participants, there is a slight difference in average pupil diameter between the two conditions (condition 2 mean = 3.224 mm; condition 3 mean = 3.199 mm), which, with greater statistical power, could suggest that mental workload was greater in condition 2, a strictly monitoring task. This would fit with the subjective ratings and the heart rate evidence. However, three participants encountered condition 2 first, with the other two completing condition 3 first. Considering the documented effects of fatigue and emotional arousal on pupil diameter, it is thus noteworthy that pupil diameter did tend to

decrease between the second and third conditions presented, regardless of which was presented first (second condition presented mean = 3.254 mm; third condition presented mean = 3.169 mm). With only five participants, the differences in pupil diameter between conditions 2 and 3, with equal luminance from the displays, are notable but inconclusive.

8.2.3 Heart Rate and Heart Rate Variability

Some of the heart rate and heart rate variability analyses produced interesting patterns. The results of other analyses did not match our expectations, and in general, this data analysis was not straight-forward. As expected, heart rate increased from the resting baseline to the first condition, and heart rate variability decreased. However, this pattern did not continue for the subsequent conditions, which were designed to be, and were subjectively reported to be, more demanding. Taken together, the differences in pupil diameter and the heart rate measures between the first condition and the following two suggest ordering effects. This could be explained as increasing fatigue or decreasing arousal during the session, as participants became more comfortable with the experiment. Pattyn et al. (2008) observed a decrease in heart rate with time on task under vigilance conditions and interpreted this as a physiological indication of underload. Whatever the reason for the effect, full studies utilizing physiological measures should train operators extensively, fully balance the order of conditions, and include multiple resting baseline periods to address these potential complicating factors.

While we did find promising results from the maximum and average heart rate measures, these are sensitive to a variety of factors. That is, a momentary spike in heart

rate could reflect heightened workload, stress, arousal, or physical activity. Based on our experience with pupil diameter and heart rate here, this is likely a challenge with physiological measures in general. Even if there are notable events in the data, it may be difficult to interpret these. However, these measures could identify moments worthy of discussion during debriefing interviews. Certainly, a full-scale study would require a larger sample for more conclusive results.

Heart rate variability is commonly regarded as more selective to mental effort effects. In general, the HRV measures seemed to indicate that participants experienced greater cognitive workload during condition 2 (monitoring four reactors) than in condition 3 (procedure-based control). This interpretation fits with the subjective NASA-TLX ratings. It is interesting, then, to note that situation awareness and periodic logging performance seem to have suffered during condition 3. This implies that workload may have indeed been lower in condition 3, due to task shedding. If this was the case, it emphasizes the need, as proposed by Guhe et al. (2005), to make a distinction between external workload (i.e., task demands) and internal workload (i.e., mental effort): the latter is subject to factors such as operator motivation, strategies and task prioritization (see also Zijlstra, 1993).

8.2.4 Relating Mental Workload Measures

We found little apparent relationship between NASA-TLX and the heart rate measures. Figure 7.19 shows some promise in relating average pupil diameter to NASA-TLX scores. For the main grouping of points, there seems to be a positive correlation, as

expected. However, there are unexplained outliers, possibly due to individual differences. The results of this analysis are inconclusive without further validation.

8.3 Situation Awareness

8.3.1 Modified SACRI

There is no redline for acceptable accuracy in the operator's present understanding of the situation. From our results, it is clear that operators are not fully aware of the system state at all times, but these numbers still seem relatively high. With an actual system, situation awareness requirements may be specified, not in terms of accuracy, but more likely in terms of specific information that is required for good decision-making.

It seems quite counter-intuitive that the response accuracy to SACRI was higher with four reactors (condition 2) than with one reactor (condition 1). There are several possible explanations for this observation. First, condition 1 was always presented first, so the participants may have still been getting accustomed to the interface and the nominal values for system parameters in this condition. At least one participant remarked that in the later conditions, he was less reliant on the tables of typical values and thresholds. It may have also been easier to guess correctly on the SACRI questionnaire when monitoring four reactors. That is, not all reactors had issues simultaneously, so the default answer of "normal" or "no change" was more frequently the correct one in condition 2. Conditions 1 and 2 were designed to have three major incidents each; the incidents averaged over four reactors means less system variability than with all incidents

concentrated within a single reactor. This potential issue is discussed further below in debriefing the methodology itself.

SACRI scores for sub-condition 3A were dramatically lower than for any other. In this part of the simulation, the operator initiated a load reduction on reactor 3, then began a load increase on reactor 1, while monitoring reactor 2 and preparing to control reactor 4. Also, reactor 4 had issues. Therefore, one possible explanation is that this scenario was complex and made maintaining situation awareness difficult. However, subjective awareness during this same period was relatively high. Based on the questions asked, it is possible that operators' mental models were insufficient for correctly predicting future developments. Some responses may have been incorrect due to varying interpretations of the "recent past," as core power in reactor 3 "overshot," descending below the new set-point and then gradually increasing to match it. However, any misinterpretation of "recent past" should have been the case in other SACRI administrations as well. Considering the problematic monitoring observed in the eye gaze data for condition 3, it is possible the operators indeed had relatively poor awareness due to a failure to monitor all four reactors as instructed. Finally, it is possible that our simplified question selection process, with all participants receiving the same questions for a particular administration, may have by chance resulted in a particularly tough combination of questions for condition 3A. Our feeling is that the phenomenon of low scores in condition 3A was likely due to a combination of more than one of these factors.

8.3.2 Subjective Situation Awareness Rating

The subjective situation awareness rating lacked sensitivity and provided relatively little insight, except for the fact that participants were reluctant to rate their own situation awareness as "Excellent" or "Poor." It seems reasonable that participants would rate their own situation awareness as higher when monitoring a single reactor than when monitoring four. It is notable that mean subjective awareness was slightly higher in condition 3 than in condition 2, meaning that participants felt at least as confident in their own situation awareness when multi-tasking between monitoring and controlling various reactors, as when they passively monitored four reactors in steady-state operation with issues. This was the case, despite the fact that we observed operators honing in on the particular reactor currently under procedure-based control, and procedure execution required significant "heads-down" time. Perhaps the goal-based, step-wise nature of the procedures gave participants a greater (and, potentially, false) sense of control and awareness of the state of the highly-automated reactors. In the second condition, with passive monitoring, participants were "side-lined" and could only watch helplessly and log and report issues as they became apparent. Multiple participants noted that this concept of operations was against their training and instinct, as they had to resist the urge to take corrective action during system incidents. Regardless of the objective level of situation awareness actually obtained, this may have been a case of "out-of-the-loop" operators (see, e.g., Endsley & Kiris, 1995), who felt they had trouble understanding what the system was doing.

It is also interesting to compare mean (normalized) self-ratings by sub-condition with the objective accuracy on the SACRI questions. The sub-condition with the lowest mean self-ratings (i.e., 2A) produced the highest SACRI scores. Also, the sub-condition with the highest mean self-ratings (i.e., 3A) consistently produced the poorest SACRI scores. That is, although there is insufficient data for a statistical argument, it would appear that these subjective and objective measures of situation awareness do not correspond very well. Previous studies have concluded that self-ratings may not be valid indicators of situation awareness, which is certainly compatible with the results here.

8.3.3 Eye Tracking

We believe the results above demonstrate the utility of high-level eye gaze data in situation awareness assessment, particularly with respect to user interface design and evaluation. Computerized analysis of the data results in a clear visual indication of the user's strategies and information sources, which can be directly applied to further iterative design activities.

The most useful analysis, which took minimal time and effort, was to look at the gaze data for individual participants in five-minute intervals, to measure conformance to the operating rules. An interesting finding was that participants distributed their attention adequately when strictly monitoring multiple reactors (i.e., in condition 2), but that some operators tended to neglect the monitoring of other reactors while controlling a certain reactor (in condition 3). Granted, such multi-tasking issues could be alleviated with in-depth training and practice, but these results demonstrate a natural tendency to

disproportionately (and, perhaps, dangerously) focus on units undergoing state changes. In condition 3, it is possible the participants became complacent about the automated reactors behaving as expected. They may have become absorbed in the control procedure because this sub-task was more engaging, to the point of losing track of time or forgetting other responsibilities. Whatever the reason, this objective measure successfully demonstrated that the operator's situation awareness could not have been as good as indicated by the self-ratings, which were highest for condition 3. This is an interesting finding: participants may have felt aware of the system, because they understood what was happening with the reactors under procedure-based control at the system goals level. However, these ratings do not reflect the fact that objectively, the operators were failing to maintain awareness of the overall system during detailed interaction with only a portion of the plant.

More detailed analysis of the eye gaze data may take much more time and effort and may produce inconclusive or misleading results. For example, perhaps the most challenging event in this experiment was when a second leak occurred on reactor 2 in condition 2. This leak was clearly indicated on the mimic display for reactor 2, but could potentially have been overlooked in all the visual noise of the eight displays. Furthermore, the symptoms for this second leak were similar to those of a pre-existing leak, meaning that changes in system parameters would not necessarily cue the operator to the new problem. While this is a highly improbable scenario, it was intended to test operator situation awareness and incident detection. Two of the five participants never

reported or logged the second leak. It is tempting to use eye tracking data to investigate operator awareness under such pre-scripted, timed events.

Figure 8.1 contains the gaze plots for three participants in the sixty seconds following the onset of the second leak indication. All three of these participants clearly fixated for a substantial portion of this minute in the area of the leak on the mimic for reactor 2. On the other hand, Figure 8.2 depicts the gaze plots for the other two participants during this minute following the leak onset. P105 (orange) does not seem to have looked at reactor 2 at all during this time-frame, and P101 (purple) only for a few moments, but not in the area of the leak.

On the basis of the gaze information alone, it seems likely that the three in the first image detected the leak, and the two in the second image could not have. However, this is not the case. P105 reported the leak within the minute, but the eye tracker frequently failed to record this operator's gaze information because she leaned forward to read the displays more comfortably. It is likely she looked at reactor 2, but that the eye tracking system failed to capture this. P101 does not seem to have been looking in that area, and indeed never reported the second leak. However, P103 clearly looked in the area of the leak on reactor 2 within the minute following its onset, but never reported it or logged it.

This example provides anecdotal evidence that eye tracking data should be interpreted cautiously with respect to situation awareness and should be supplemented with other information, such as operator actions or verbalizations. An operator may look at a display element and process it and understand it. He or she may also fail to recognize

it or understand it, despite looking at it. It is also possible that an operator sees and detects something, but that the eye tracker fails to record these moments. Of course, the operator may never look at the information, and thus fail to detect it, but it is hard to tell the difference in post analysis without totally reliable data. Finally, as noted by Endsley (1995), eye tracking fails to reflect the perception of information through peripheral vision.

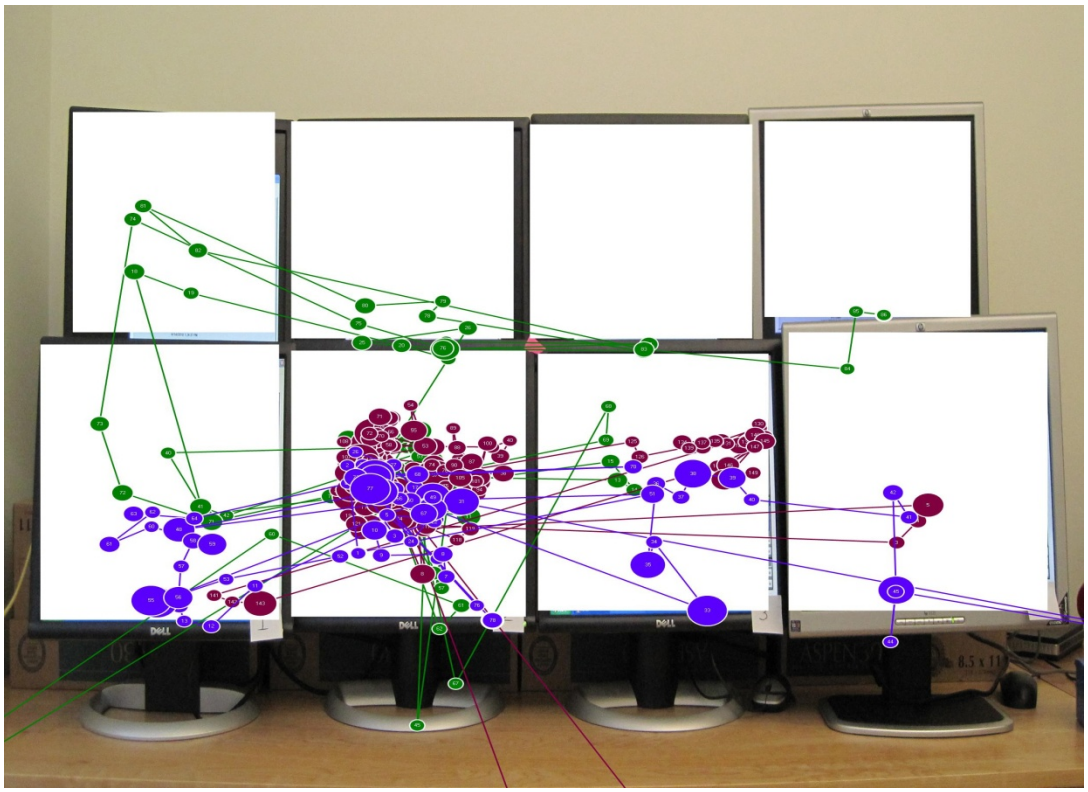

Figure 8.1. Eye Gaze during Onset of Second Leak on Reactor 2 for P102, P103, P106.

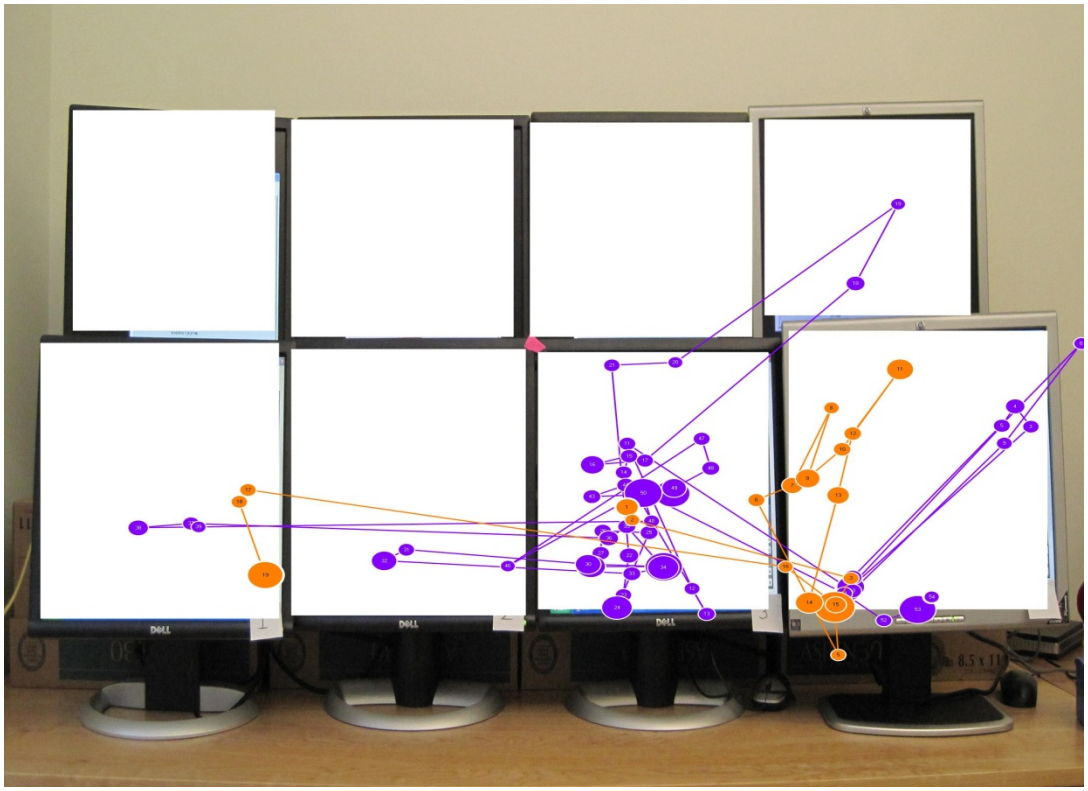

Figure 8.2. Eye Gaze during Onset of Second Leak on Reactor 2 for P101, P105.

8.4 Relating Mental Workload and Situation Awareness

This experiment produced a substantial range of individual NASA-TLX ratings (from 8.3 in condition 1C to 83.3 in condition 2B, both for participant P101). Unfortunately, individual participant situation awareness scores, either subjective or objective, were much less sensitive to the task manipulations. SACRI scores ranged from 7.5 to 12 correct responses out of 12 queries, but most scores fell in the 9 to 11 range. Some studies have found a relationship between situation awareness and mental workload for certain cases, while others have concluded these are largely independent, and may appear in various combinations. Figure 8.3 presents a scatter plot of the composite NASA-TLX score and the number of correct SACRI queries at each freeze or

stoppage of the simulation (i.e., one data point per sub-condition per participant). Based on this data, the two constructs appear to have had little relationship in this experiment.

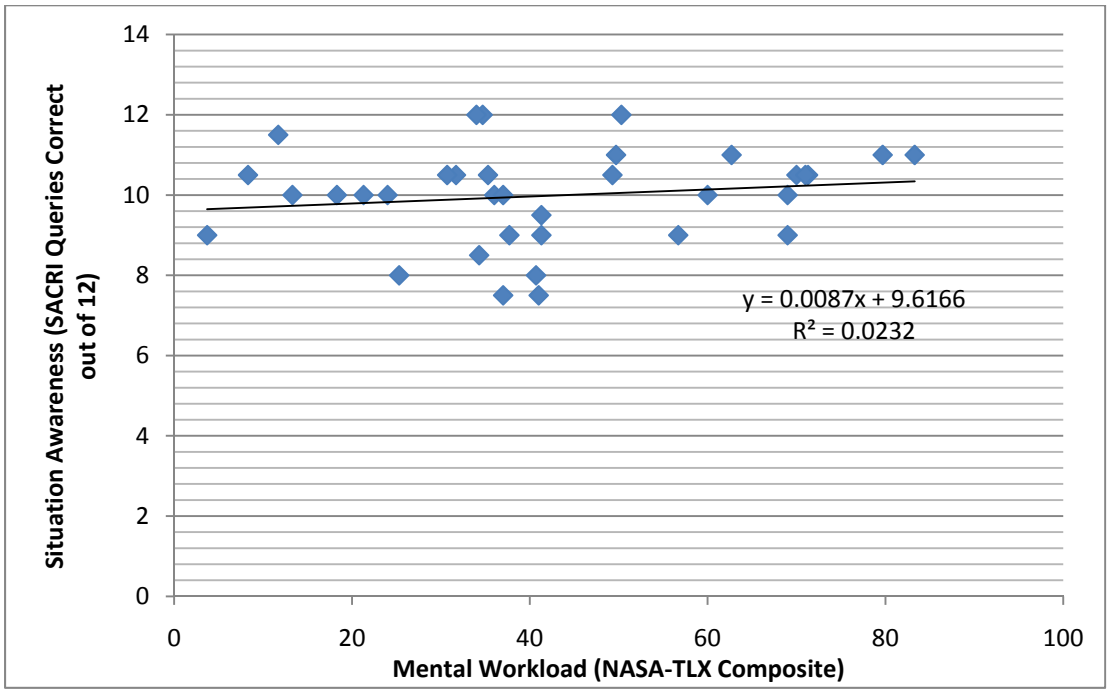

Figure 8.3. SACRI x NASA-TLX across all Sub-Conditions and Participants.

However, it is interesting to compare the results averaged across sub-conditions and participants (refer to Figures 7.4, 7.21 and 7.22). On average, mental workload was subjectively rated as highest in condition 2, in which the participant had to monitor four reactors and report incidents, while the mean subjective situation awareness score was slightly lower for this condition than for the other two. Subjectively, then, participants had to work harder to monitor the four reactors and felt less confident in the resulting situation awareness. However, the objective SACRI scores were the highest for this condition, on average, meaning that objectively, the additional effort expended seemed to maintain relatively high situation awareness. It seems counter-intuitive that objective

situation awareness was higher with four reactors than with only one, but this issue has been discussed above.

Based on our experience with physiological measures of workload here and their selectivity issues, we regard the prospect of applying these to situation awareness measurement with serious reservations. We can see how situation awareness could be assessed by capturing stimulus detection events physiologically (e.g., in EEG, heart rate, or galvanic skin response data), as discussed by Wilson (2000). However, we promote eye gaze as the most reasonable physiological measure of situation awareness available at present.

8.5 Debriefing of Methodology

The primary purpose of the pilot experiment was to serve as a proof of concept for the measures recommended at the intermediate budget level. We observed variation in the measure values between conditions, which, assuming the validity and proper application of the various techniques, makes a case for their sensitivity to the experimental manipulations in our simulated environment. However, due to the small sample size and lack of statistical power, we cannot present our results with confidence as evidence for or against the validity or sensitivity of the various measures investigated. This is especially the case as some quantitative results did not fully fit with our initial hypotheses. It is easy to find apparent counterexamples to the validity of the measures based on our results. However, arguing against their validity is not so straight-forward, as such instances can be explained not only by validity issues with the measures, but also by incorrect hypotheses, limitations of the experimental design or problematic application of

the measures. For the most part, we did not find evidence of relationships between the various measures (e.g., HRV and NASA-TLX, subjective situation awareness and SACRI, NASA-TLX and SACRI), but this seems to be in line with previous work (e.g., Endsley, 1993; Salmon et al., 2009).

There are a few notable exceptions, however. Figure 7.19 provides hope that careful investigation with a wide range of task difficulty may demonstrate positive correlation between pupil diameter and NASA-TLX (i.e., physiological and subjective measures of mental workload). In the gaze analyses of situation awareness, the visual presentations of the results demonstrate good face validity: for example, the composite heat maps look the way one might expect for each experimental condition. Also, the alignment of heart rate and pupil diameter data for participant P105 in condition 1A, particularly as these correspond to her behavior and subjective reports, provides support for the two measures in terms of convergent validity. However, in this instance they appear to have reflected psychological stress more than mental workload, per se. The periodic log performance and gaze analyses both seem to reflect some task manipulations, as well as some differences between participants related to previous experience. NASA-TLX ratings per participant appeared fairly stable over repeated administrations within a given condition. At the very least, these observations provide circumstantial evidence supporting the validity of the measures.

The majority of the remaining discussion regarding the measures deals with other practical considerations from our experience and the reports of the participants (i.e., we focus on new insights in terms of the five criteria for measure selection besides validity

and sensitivity). We endorse those measures for which we did not identify major unexpected issues in terms of these other criteria (i.e., our experience was favorable), and flag those measures which we found problematic in practice. This section concludes with revised ratings of the techniques applied in the experiment in terms of the seven criteria.

8.5.1 NASA-TLX

NASA-TLX was extremely simple to administer and analyze, even using a pen-and-paper format and manual data entry into a spreadsheet for analyzing the results. A computerized version would further streamline data collection and analysis. The composite scores appeared to be sensitive to the condition manipulations, but there were fairly large individual differences. Administration typically took less than a minute. On average, participants rated the instrument a 7.3 out of 10 in its ability to capture their level of workload, stress or effort. One participant (P101) remarked that he felt his responses were somewhat arbitrary and variable, but his composite scores were in fact fairly consistent within each condition.

Because of our positive experience with NASA-TLX here and its wide adoption, we certainly recommend NASA-TLX as a subjective measure of workload. However, each dimension should be carefully explained prior to the session, and the participant should be calibrated for what the extremes of the bipolar scales mean (see, e.g., Colle & Reid, 1998). For example, “very high” physical demand is subjective in the undesirable sense (i.e., it is open to interpretation). However, an illustrative example of “very high” physical demand from everyday experience or the participant’s previous experience may

help to calibrate the responses and reduce inter-individual variability. One other slight issue we found with NASA-TLX is that in the sub-dimension weighting procedure, it is unclear how to assess how much "own performance" contributed to workload, as opposed to, e.g., mental demand.

8.5.2 Modified SACRI

Overall, participants responded to the SACRI questions with relatively high accuracy across conditions. The one exception was freeze 3A, in the middle of condition 3: scores were consistently lower for this administration. This could reflect lower situation awareness in the minutes prior to the freeze or a combination of other factors, as discussed above. From this study, it is not clear how sensitive a simple response accuracy score would be to manipulations in monitoring tasks. Endsley et al. (2000) believed that this simple scoring is not appropriate for SACRI's kin, SAGAT. Also, the developers of SACRI incorporated Signal Detection Theory in scoring the responses (Hogg et al., 1995). This approach is more labor-intensive, but is probably warranted.

We also did not fully follow the SACRI method in that freeze times were consistent across participants, and at a particular freeze, each participant received the same randomized set of questions. This may be acceptable for a pilot study, but even this simpler approach with pen-and-paper administration required significant effort for question generation and inventory scoring. If SACRI is adopted for efforts of larger scope, it is likely worth the effort to develop an application for administering SACRI. Scoring would include instrumenting the system to capture relevant parameter values at

the appropriate times, but would probably still require subject matter expert judgments in some cases (Endsley, 1995).

Therefore, whether administered via paper or computer, proper administration of SACRI (or SAGAT) requires significantly more effort than a subjective rating of situation awareness (e.g., SART) or mental workload (e.g., NASA-TLX). However, operator acceptance was relatively high: most participants felt that the questions were fitting and captured their level of situation awareness well. Of the two Navy reactor operators, P106 rated SACRI at 10 out of 10 for fitting with his mental model, and P101 gave the measure an 8.5. Regarding the question content, a participant with TRIGA operating experience and industrial power plant experience (P102) said "I think that's pretty much what an operator would monitor." P106 rated SACRI a 9.5 out of 10 for capturing his situation awareness, explaining that the questions "were good… those are the kind of questions you would ask the operator and they should be able to know [the answers]."

The simulator freezes for SACRI administration presented some minor problems. During each simulator freeze, we first took the opportunity to administer NASA-TLX, and, at the end, we re-calibrated the eye tracking prior to resuming the simulation. Altogether, the shortest freeze lasted 3:22 and the longest was 7:01, with duration of the twenty freezes averaging 5:09. Four of the five participants reported that, at some point, a freeze was disruptive of the task. Freezes caused operators to temporarily lose their place in procedures (P102, P105, P106) or to forget their place in the logging sub-task (P101). However, most felt the freezes were "minimally" (P105) or "not very" (P106) annoying.

On a scale from 1 to 10 (very annoying), P101 rated the freezes as 3.5 and P103 rated them as 7. P102 said "no," they were not annoying. P101 and P105 reported that the freezes provided a mental break from the tasks, "a moment to pause and think" (P101). In some evaluations this is probably undesirable. The end of a SACRI freeze provided an opportune time for re-calibrating the eye tracker, but it also sometimes resulted in temporarily heightened heart rates upon the resumption of the simulation. One of the most challenging aspects of administering the experiment was to remember to mark each freeze and unfreeze as a "lap" for the heart rate recording. Subsequent implementations should also include a mechanism for instantaneously freezing the simulator and blanking the screens. In our setup, there was a lag of a few seconds for manually freezing the four simulators prior to blanking the screens. At least one participant viewed the screens intently during this time, which could have boosted his performance on SACRI.

We attempted to adapt SACRI for a control room with multiple reactor modules with mixed results. Participants reported that in some cases, it was difficult to "keep everything straight," in order to answer the SACRI questions correctly for each unit, which met with expectations. However, varying the number of operating reactors had unforeseen interaction with the SACRI measure of situation awareness. It seems likely that maintaining situation awareness with four times the parameters is more difficult, as reported by the participants, but on average, it may have been easier to guess the correct answer with four reactors than with one. In reasonable scenarios with fully independent reactor modules, it is highly improbable that more than one would have a major issue at the same time. Therefore, the null hypothesis (e.g., a parameter value is "normal" or

stable) is likely the correct response and thus is a very safe guess. In other words, assuming not everything is going poorly, the answer to many questions can be guessed correctly, regardless of the actual level of situation awareness. In terms of Signal Detection Theory, with randomized questions across all operating reactors, one would expect most questions to result in "correct rejection." Therefore, a relatively large number of dynamic events or parameter fluctuations would need to be scripted to produce useful results (i.e., an estimate of hit rate). As pointed out by Patrick et al. (2006), such a randomized, global approach may not capture the aspects of situation awareness that are critical during a particular manipulation. This is particularly true in a multi-modular reactor control room.

In administering SACRI and scoring the responses, the largest issue we encountered was related to wording. We provided the operator with a table of the parameters to monitor and a typical value for each. Instructions specified that "normal" for a parameter is "close to this value," but this still left room for subjectivity. For example, P101 wanted to qualify his responses to the multiple choice questions with modifiers (e.g., decreased "slightly"). SACRI and SAGAT are presented as objective measures, so there needs to be a clear correct answer. When generating the answer key, we lacked the domain knowledge or simulator experience to confidently select the "correct" answer to some questions. That is, both operators and experimenters need to objectively agree on what "normal," "recent past," "near future," "increase" and "decrease" mean. Otherwise, the operator may have good awareness but judge the situation differently than expected.

Following Hogg et al. (1995), we operationally defined the recent past as three minutes prior and the near future as three minutes beyond the current time. However, we failed to communicate this definition to some participants, resulting in cases where the participant apparently used a longer time window for the "recent past." Therefore, either experimenters need to be more specific in their instructions and expectations, or the questions should be re-worded to incorporate these definitions explicitly. An objective definition of "normal," as advocated by Endsley (1995), is not straight-forward for a new system and freshly trained test operators. Unless some rule of thumb, such as a tolerance of ±1%, is used across the board, some system expert must specify a tolerance for each process parameter, and then the operator must learn all of these prior to data collection.

In conclusion, we believe SACRI is a reasonable measure of situation awareness. However, based on the effort required for preparation and analysis and the somewhat undesirable need for freezes, we are less confident in recommending this measure following our first-hand experience. Its utility depends on the types of research questions faced by a human factors program; we believe it would be more useful with an actual system and experts who can specify clear, prioritized situation awareness requirements for a given scenario. However, without changing the rules of question generation, there is no guarantee these specific requirements will be tested (Patrick et al., 2006). If SACRI is to be adopted for a multi-modular reactor control room, further thought and adaptation is required to address the problems discussed above.

8.5.3 Eye Tracking: Pupil Diameter and Gaze

With the Tobii X120 and the Tobii Studio software, pupil diameter recording is fairly easy. At 60 Hz, however, the high-volume data quickly becomes cumbersome. Because the task-evoked pupillary response is momentary, there may be more effective analyses than averaging over intervals (Iqbal et al., 2004), but these require significantly more effort than our time-averaged method, either by hand or by developing automated analyses.

Because our application of the pupil diameter measure was problematic, we cannot fully answer whether this technique is feasible for the target environment. However, our results and a further review of the literature suggest that pupil diameter is not particularly selective to cognitive load. That is, pupil diameter may be sensitive to mental effort, but fatigue, arousal and, obviously, illumination levels also affect this measure. These may be difficult to control experimentally in a naturalistic simulator setting. Pupil diameter may thus be a risky choice for measuring mental workload in similar studies of long-term monitoring tasks. However, it may be a useful general indicator of operator fatigue or arousal, if display luminosity can be carefully controlled. In this study, the information presented on each monitor was fixed. Controlling luminosity may be more difficult if the human-system interaction involves navigation to various display pages. If pupil diameter is used as a measure of mental workload, the order of conditions must certainly be balanced. In any experiment using physiological measures, resting baselines should be obtained, at minimum, at the beginning and at the

end of the session, to help identify other factors contributing to measure variation such as fatigue or boredom (see, e.g., Healey & Picard, 2005).

As long as a remote eye tracking system is compatible with the environment and tasks, we recommend gaze-based analysis of situation awareness. We encountered limitations with our eight-flatscreen display setup related to the physical specifications of our system. In order to limit the field of view to reliably track gaze to the four corners of the display area, we had to seat the operator at a larger-than-optimal (and perhaps slightly unrealistic) distance. At this distance, the geometry of the displays sufficiently resembled a plane such that, despite some distortion due to parallax error, we could determine what general area of which screen was attended at a particular time. However, larger or more complex display geometries would pose practical challenges in eye tracker configuration. For these reasons, Hornof et al. (2010) have resorted to simulating focal and peripheral vision on a single display for investigating visual attention in multi-display configurations with a remote eye tracker. This is a labor-intensive but elegant solution.

On average, participants felt the eye tracking was only slightly more distracting during their tasks than audio/video recording (2.0 out of 10 vs. 1.25 out of 10, with 10 being “extremely distracting”). P101 noted that the eye tracking limited him from “scooting around” in his chair. Overall, though, except for the calibration procedure, this measure is highly unobtrusive. If periodic freezes are used for SACRI, re-calibrating prior to resuming the simulation is a good solution. Alternatively, Hornof and Halverson (2002) developed custom software to automatically determine when re-calibration is necessary, based on predictable fixations. It seemed to us that during the course of a

typical session, the calibration process gradually became more smooth and efficient. If this was indeed the case, we are uncertain if the system adapts to the individual over time or if the individual was simply learning the ideal body position for calibration.

Basic gaze data analyses can be performed quickly and, as demonstrated above, can address high-level design and evaluation questions directly. The analysis software supports experimenter-specified time and participant filters, so any interval of gaze data can be easily inspected on a collective or individual basis. As noted in the literature (e.g., Endsley, 1995), gaze does not imply understanding, so gaze should not be the only measure of situation awareness. However, because of the rich, intuitive nature of the resulting visualizations, this technique can efficiently identify potential problems in the user-interface-task system. That is, with reliable recordings, it is very clear what information sources the operator neglected, suggesting weak spots in awareness. Such findings could directly serve as input into subsequent training, task and user interface design efforts.

8.5.4 Heart Rate Monitoring: Heart Rate and HRV

Heart rate data collection with the Polar WearLink/Transmitter and the Polar RS800CX training computer was easy, unobtrusive and relatively inexpensive. In the post-session interviews, all five participants responded that the chest band never got in the way of their tasks. Three stated that it was never uncomfortable; one reported discomfort only when walking around the building during breaks, and the remaining participant stating that he only *noticed* the chest band's presence during the breaks. This

all implies exceptionally high operator acceptance and low task intrusion for a physiological measure. In the total of approximately twelve hours of recording, we only failed to obtain the wireless signal for a couple brief periods. The wireless receiver wristwatch, controlled by an experiment assistant, was located several feet from the operator's workstation and could have been moved further away without impeding connectivity. If the R-R recording feature indeed provides accuracy comparable to EKG, as evidenced by an independent validation (Nunan et al., 2009), then this wireless chest band approach seems to be a step forward for human factors studies, considering the cost and practical issues of traditional EKG recording.

Data analysis using the Polar software was also straight-forward. For a given recording, any desired portion could be highlighted and automatically analyzed for average heart rate and various measures of heart rate variability. Also, erroneous spikes in the heart rate data could be automatically identified and eliminated at a user-specified level of tolerance. There were typically only a couple such errors in the data per condition per participant; many recordings were error-free. During data collection, we used a feature of the wristwatch computer to mark each freeze and unfreeze of the simulator as a "lap," which then proved useful in data analysis.

We did not conduct a quantitative analysis of the effect of freezes on heart rate data. However, visually inspecting the plots, there does not seem to be a consistent effect of freezes and resumptions on heart rate. For some participants in some cases, the heart rate temporarily elevated upon resumption after a freeze. Some participants reported that the freezes provided a mental break, and this was reflected in the heart rate data for the

more experienced operators in condition 1. This trend was less obvious in subsequent conditions. The heart rate varies to a surprising degree throughout a scenario, and the freeze intervals generally fit this overall pattern of variation.

As with the pupil data, a major challenge is the volume of the data. The freezes were included in the heart rate recordings, which made analysis somewhat more tedious. Compared to subjective measures, the physiological data is not "clean," and is therefore more difficult to work with. For this study, the lack of reliance on statistics limits the claims that can be made, but some of the trends in the heart rate and variability data seem to match with subjective reports of workload and stress. We found the results of this technique compelling enough to believe that heart rate recording with the more convenient wireless equipment is feasible and indeed shows promise as a physiological measure of workload. However, even if statistically significant differences are found between conditions, it may be difficult to interpret the results, due to the limited ability to distinguish between workload, stress, fatigue, arousal, and other factors. That is, if strictly workload is to be measured, selectivity may be an issue, at least for average heart rate. Finally, the number of heart rate variability measures available may pose challenges, as the experimenter must select a subset of these for analysis. If the measures identified produce conflicting results, interpreting these differences may be difficult.

Both heart rate and heart rate variability have been fairly widely used. Our impression from the literature is that there is greater agreement in the field on the variability measures as valid workload indicators, as compared to heart rate itself. This was reflected in our initial matrix-based scoring. However, the moment-to-moment heart

rate data is available with this technique at essentially no additional cost beyond that of measuring HRV. Therefore we believe that heart rate might as well be included in the methodology and analyses, despite its questionable validity. That is, it may provide a useful perspective on the overall demands placed on the operator at little extra cost. If average heart rate data is presented as an indicator of mental workload, though, it should take a secondary role to the other measures applied, and a caveat should be made regarding the validity question.

8.5.5 Revised Criteria-Based Ratings

In Chapter 5, each measure under consideration was numerically rated based on the seven composite criteria for measure selection. Based on the discussion above, we have revised the criteria-based ratings for those measures applied in the pilot study. The results of the finalized ratings are summarized in Table 8.1. The criteria weightings for the weighted score remain the same. Cells containing revised ratings are emphasized with (+) if our judgment increased in favorability and with (-) if we have greater or new-found concerns based on our experience.

NASA-TLX was easy to administer and analyze. Pupil diameter exhibited selectivity issues which will likely pose challenges to application of this measure in a complex simulated environment. The wireless chest band approach to heart rate monitoring was highly favorable as it did not disrupt operator tasks and no one reported discomfort during the experiment. However, various measures of heart rate variability may produce contrasting results, such that interpretation is not straight-forward.

Even with our simple approach, SACRI required more effort than we expected. A software application for administration would eliminate some, but not all, of these hurdles. The SACRI measure generally did not vary much between administrations or participants. We have also discussed various issues with the need for simulator freezes, including the potential for interference with other measures (i.e., relatively low compatibility). Finally, a remote eye tracking system requires a significant up-front investment, but the types of gaze analyses we performed required less effort than we expected.

Table 8.1. Final Criteria-Based Ratings of the Mental Workload and Situation Awareness Measures Applied, with Changes Emphasized.

	Sensitivity/ Accuracy/ Precision	Validity/ Repeatability/ Reliability/ Selectivity	Unobtrusiveness/ Operator Acceptance	Compatibility with Prob. Constraints	Level of Adoption/ Consensus	Interpretability/ Diagnosticity/ Redlines	Convenience/ Cost to Implement	Total Score (Wtd.)	Total Score (Raw)
MW									
NASA-TLX	4	3	3	4	4	2	4 (+)	51	24
Pupil diameter	3	2 (-)	4	2	3	2	1	38	17
Heart rate	3	1	4 (+)	3	2	1	3	35	17
HRV	3	3	4 (+)	2	3	1 (-)	3	41	19
SA									
SACRI	2 (-)	3	2	2 (-)	2	3	1 (-)	34	15
Gaze	2	3	4	2	1	4	1 (+)	38	17
WEIGHTS	**3**	**3**	**2**	**2**	**2**	**2**	**1**		

8.6 Implications for Human Factors in Monitoring Multiple, Highly-Automated Reactor Modules

As expected, subjective workload was higher for the conditions in which the operator was assigned to four reactor modules than to only one. However, it appears that ordering effects impaired our ability to compare these conditions from a physiological standpoint. Two participants (P102, P106) reported that the difficulty of the four-reactor monitoring task (condition 2) was approximately twice that with one module (condition 1). Despite some issues with our adapted version of SACRI, it appears that operators were able to maintain fairly high situation awareness with four reactors in operation, at the cost of considerable workload. It is not clear how this finding from a short-term monitoring session with minimal training would compare to the workload and situation awareness over long shifts. We did find that despite their fixed spatial locations and numerical ordering, referring to reactors by number sometimes caused confusion and inaccurate statements, even among the experimenters.

Participant views and strategies within this concept of operations were interesting. We provided a scan pattern for ensuring coverage of the parameters within the mimic display for a single reactor, and instructed the operators to loop through the four reactors from left to right (i.e., numerical order). Several reported that it made more sense or was easier to check a single reading at a time across all four reactor units. This approach is reasonable during normal operations, and design efforts for multiple units should consider a display view which groups related parameters across units for quick verification. However, such a piecemeal scan strategy or display may hinder proper

formulation of the correct situation model per unit, since process parameters are inter-related, and combined, these describe higher-level status information.

In some respects, the control room simulation conflicted with the expectations and prior training of those participants with operations experience. When problems arose, they reported that they had to fight their instincts to take corrective action immediately. It was counter-intuitive to leave control up to the automated systems and simply observe. They felt the five-minute logging period was burdensome. Our simple alarm system failed to provide the information they desired, and the system did not respond to cases where a parameter crossed a threshold in the way they expected. These examples show that using operators with previous experience on a dissimilar system has benefits and drawbacks: they provide helpful domain knowledge and insight from experience, but their actions and opinions may be influenced by prior training.

It is noteworthy that in comparing conditions 2 (monitoring four reactors) and 3 (control and monitoring), participants reported lower workload and slightly higher situation awareness in condition 3, but in this condition, they performed more poorly on the logs, failed to adequately monitor all reactors, and objectively showed poorer situation awareness. We designed these two conditions with the hypothesis that workload would be higher when performing procedures, but that situation awareness would be aided by procedures keeping the operators "in the loop." In reality, operators seem to have focused on the reactors undergoing evolutions to the point of neglecting others. Whether this task shedding was conscious or not, it served to moderate workload, but this was at the cost of reduced situation awareness overall. In general, the operators appeared

quite engaged in the procedures, but some exhibited poor vigilance for the reactors in steady-state operation. The understanding of the system state and goals afforded by the procedures may have lulled operators into a false sense of situation awareness: even if they attained better awareness by being kept in the loop for the reactors under control, they were observed to neglect the rest. On the other hand, the operators had to work hard to maintain awareness during passive monitoring of the four reactors, but they were apparently successful.

Dixon and Wickens (2003) found some evidence of cognitive tunneling during the performance of difficult UAV control and monitoring tasks, which resulted in task shedding. This is compatible with our observations in the third condition: such tunneling may occur due to task complexity during normal operations, not just in high-stress conditions.

It is not clear how the actual numbers obtained in our simulated tasks would compare to those in a high-fidelity simulation of an actual plant. There would potentially be many more parameters to monitor in the actual system, but training and interface design would be significantly more involved as well.

8.7 Design Implications

This experiment demonstrated the potential for operators to become engrossed with the actions of one reactor to the point of neglecting monitoring duties for the others. That is, attentional tunneling is a potential issue. Tunneling is a commonly cited problem under stressful conditions, but participants in this study exhibited similar behavior during execution of normal control procedures. These conditions were demanding but not

stressful, according to most participants. Thus the interface design should limit either the depth or the consecutive time devoted to interaction with a single unit, for both normal and abnormal conditions. Potentially, the user interface could limit interactions to small time chunks, providing natural breaks for resumption of normal multi-modular monitoring. Operating rules may address this tendency too, but if the problem is distortion of time (see, e.g., Hancock & Szalma, 2003), relying on operator conformance to time-based rules may be ineffective. We therefore believe the interaction design for computerized procedure-based control is particularly important. Even simple procedures may need to explicitly guide the operator, not just to task completion with the associated reactor, but to maintain awareness of the other reactors. For example, in our experiment, the procedures included steps of assessing the status of the other reactors assigned to the operator.

The crew allocation model is another aspect of designing to address the potential issue of attentional tunneling. One participant (P101) remarked that when abnormal conditions are detected, he would expect that second crew member to be assigned to assist, such that one works with the problematic unit, and the other continues the normal monitoring task for the remaining units.

In our experiment, periodic logging on paper was intended to keep operators engaged in the monitoring task, to encourage situation awareness, and to provide a straight-forward basis for judging task performance. However, the frequency of periodic logging was objectionable to some participants. In modern systems, paper logs may be eliminated, as automated logging is a simple and reliable solution. However, a

completely passive monitoring task may be objectionable to operators. Therefore some periodic human-system interaction is advisable, even if no manual control is presently required. A digital log interface could periodically pop up on an interface screen, requesting momentary attention from the operator. This approach would address the issues of neglected logs seen in our study under complex multi-tasking. Another alternative is a wizard-type interaction paradigm, in which, in addition to normal display monitoring, system software periodically leads the operator through key parameters, requiring operator sign-off in order to continue. If logging is fully automated, it may be desirable to support secondary notation (Blackwell & Green, 1999) on, for example, archived trend plots. Operators could be allowed to add "sticky-note"-type comments to noteworthy events in the automatically logged data.

We found it interesting that some operators preferred to scan a single parameter across all four reactors (e.g., pressure on reactor 1, pressure on reactor 2 …), rather than scanning all parameters on one reactor before moving to the next reactor (reactor 1 power, pressure, temperature, etc., reactor 2 power, pressure, temperature, etc. …). The former strategy may be more efficient, less mentally demanding, or both. This finding suggests there are two types of monitoring tasks which contribute to situation awareness at different levels and answer somewhat different questions. A user interface could provide different views of the data in support of both tasks. First, grouping parameters by similarity (e.g., core power across all four modules) supports a check reading task, in support of quickly assessing whether the system is operating within normal limits. This implies efficient attainment of Level 1 situation awareness. Second, grouping parameters

by system proximity (e.g., all core sensor data for reactor 2) supports awareness at the subsystem and reactor level. The second type of monitoring may require more effort for attaining basic system awareness, but this effort likely produces superior situation awareness at Levels 2 and 3, as the operator can better integrate related information per reactor into a composite picture, and predict future system behavior. In other words, the check readings task across all four reactors does not necessarily result in a deep, holistic view of the state of a particular reactor.

Vicente and Rasmussen (1992) proposed a levels of detail-based presentation of complex system information, and this approach seems sensible to us. We foresee a design approach in which the default, "zoomed out" view provides high-level awareness of system status. If this information is presented in a mimic (i.e., simplified schematic) view of the system and processes, the interaction design could feature high spatial compatibility of the displays and controls: for a more detailed view of a subsystem, the operator need only select it directly on the display.

Finally, color may be used in various ways. However, with a complex system interface, color will be most meaningful if used sparingly and consistently to encode information. Within the levels-of-detail design approach, color could be used to logically link key information which is spatially distributed across one or multiple displays. For example, if there is an alarm due to a high temperature reading on a particular sensor, the location of this sensor, as well as the alarm message, could be highlighted in matching colors. This design approach is a direct application of information foraging theory,

providing perceptual cues for information scent. Clearly, this approach would require a small color palette and muted color for normal information presentation.

8.8 Proposal of Task-Interface Characterization Rubric

One final research objective remains to be addressed. Because there are multiple indicators of system-task-interface design adequacy, and multiple such techniques may be used to evaluate a system, we believe a standardized approach for organizing and presenting this multi-dimensional data would be useful. For example, in support of design certification by the NRC, the proposed design of a nuclear power plant control room itself should be accompanied by empirical measures of design adequacy, such as mental workload, situation awareness, performance metrics, and related probabilistic risk analyses. If this data is available on a per-task basis, it is appealing to produce a composite snapshot or profile of the system, as characterized along these dimensions. This could serve as a rubric for human-system interface evaluation on a task-by-task basis.

One important consideration is the selection of tasks for evaluation. For complex systems, it is not feasible to test every foreseeable condition, task and user interface screen in the lab or simulator. Therefore, a representative sample should be chosen (see, e.g., Ha et al., 2007; Norros & Nuutinen, 2005; O'Hara et al., 2004, Section 11.4.1). For an evaluation to be summarized by a rubric, a stratified random approach to selection should be employed, which ensures inclusion of tasks performed frequently under normal conditions, including some approaching automaticity of performance; less frequent,

procedure-driven tasks; and rarely encountered emergency response tasks. Also, for staffing exemption requests, NUREG-1791 emphasizes the importance of considering "the operational conditions which present the greatest potential challenges to the effective and safe performance of control personnel" (Persensky et al., 2005, p. II-3-1).

The goal is to devise an instrument or rubric that is light-weight and prescriptive. That is, the effort should mainly be focused on the requisite system evaluation itself, with the primary cost of the rubric-based evaluation being the time and effort of compiling the multi-dimensional data into a single format. Also, the method of presentation should help to identify problematic, outlier tasks, for which procedure or user interface re-design is recommended. The resulting data format should represent the system in a manner which supports judgments about the adequacy of the design overall, as well as comparison of results between competing prototypes, similar systems, or even across domains.

The various dimensions (e.g., workload, situation awareness, performance, error rate and impact) provide differing perspectives on design adequacy, but they are inter-related. For example, Endsley (2000) noted that situation awareness and performance are probabilistically related. Similarly, for two tasks with comparable performance in the simulator, if one requires significantly higher mental workload, it may be more problematic because it is more vulnerable to a performance decrement in actual operations.

Regarding underload effects on performance, recall the two competing models (O'Donnell & Eggemeier, 1986, and the inverted U-curve model, e.g., Proctor & Van Zandt, 2008). Perhaps neither of these quite captures the potential problems of underload

accurately. Multi-tasking performance must necessarily decrease in the overload region. On the other hand, performance may or may not suffer due to underload conditions (e.g., Durso et al., 1999). We speculate that the inverted-U curve of performance by level of mental workload may therefore really mean that the *potential* lower boundary for performance is worse for underload conditions compared to intermediate workload, not *necessarily* that performance will suffer with low workload. Therefore, this curve is perhaps over-simplified, in that it expresses two somewhat different phenomena at the workload extremes.

Kaber and Mosaly (in Grier et al., 2008) discussed a multi-dimensional redline for system design adequacy, incorporating workload, situation awareness and other dimensions. In proposing a rubric here, we agree that redlines may make most sense from this multi-dimensional perspective. For example, if error impact is high *and* mental workload is high, then the evaluator can more confidently note an issue. A complicating factor is that duration may also be a factor. For example, long-term high or very low workload may result in performance decrements not observed in brief scenarios.

Endsley (1993) plotted individual estimates of situation awareness and mental workload together. This is particularly useful when analyzed in the context of Endsley's (1993) two-dimensional theoretical plot which relates the two constructs by region. Although there may be little or no correlation over the data set, the individual points themselves, and taken collectively, are meaningful indicators of task and user interface design in terms of system safety. That is, soft redlines could be constructed within such a plot, identifying problematic points. Redlines could even be specified at graduated levels,

rather than a hard cut-off boundary. Such regions would help to identify those tasks most urgently in need of interface re-design.

The clear presentation of information becomes challenging with more than two dimensions. Hallbert (1997) included a three-dimensional plot summarizing his results, which added team interaction to the constructs discussed above. This approach illustrates some of the conclusions of the study nicely, but is, overall, difficult to read.

We briefly present two notional approaches here, serving as a first step toward a task-interface characterization rubric and, hopefully, conveying the usefulness of the concept for iterative system design and evaluation. We then leave further thought and validation for future work.

One possible approach for combining multiple dimensions of information is a table-based approach. With this approach, each row summarizes the evaluation results for a single task or procedure considered. The columns capture the composite scores for, e.g., mental workload, situation awareness, and performance metrics. It would be possible to devise an overall design adequacy score per task, based perhaps on a product or weighted sum of the preceding columns (Table 8.2), with low adequacy scores indicating opportunities for re-design. Without empirical validation, however, such a formula may be somewhat arbitrary. Selection of representative tasks could be based on first having subject matter experts select procedure/task combinations at various anticipated levels of overall demand or workload, from potential underload to potential overload.

Table 8.2. Table-Based Approach to Multi-Dimensional Task-Interface Evaluation Rubric.

Task Name / Proc. No.	Mental Workload (NASA-TLX)	Situation Awareness (SACRI)	Performance (Efficiency)	Design Adequacy Score
Reactor Startup (GOI 5-1)				
Load Increase (GOI 5-4)				
Reactor Shutdown (GOI 6-1)				
Loss of Coolant (EOP 3-7)				
Alarm Response (ARP 5-12)				

The second approach, which harnesses the strengths of human visual perception, is an information visualization-based approach. The mockup in Figure 8.4 follows the precedent in the literature of plotting situation awareness on the y-axis and mental workload on the x-axis. Endsley (1993) discussed the various regions of this two-dimensional continuum, but did not account for the issues of underload. Obviously, high workload and low situation awareness is an undesirable combination. Considering the potential ill effects of underload, intermediate workload and high situation awareness may be optimal (Alexander et al., 2000). The light blue arc in the figure emphasizes this theoretically optimal region. For illustrative purposes, we add two additional dimensions, error rate and error impact, encoded by color and size, respectively. Integrating the four dimensions, even tasks which are characterized by high situation awareness, but which contain high workload and high error impact are undesirable. A four-dimensional redline is not straight-forward, but outliers requiring further design efforts should be apparent.

Also, the overall snapshot of the system by clusters of representative tasks gives a summary of the findings of human factors evaluation activities, and provides a basis for comparing rubric signatures between system designs or even across domains. Outlying tasks may stand out perceptually, as opposed to the reliance on a formula for calculating an overall adequacy score above. We are not set on these four particular dimensions, but the visualization-based approach exemplified here, although perhaps less formal, appears to us to be more usable than the table-based approach.

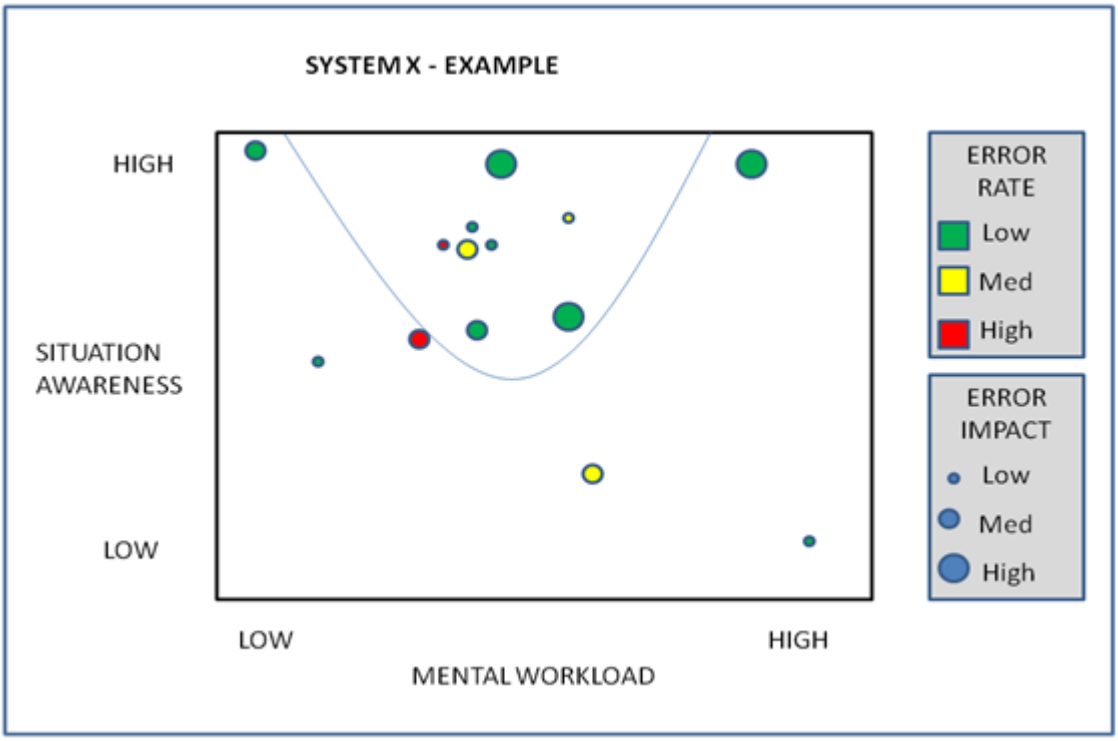

Figure 8.4. Proposed Information Visualization – Based Approach to Multi-Dimensional Task-Interface Evaluation Rubric. High situation awareness and intermediate workload are assumed to be optimal.

9. CONCLUSION

NuScale Power's Human Factors Engineering program needs to assess the levels of mental workload and situation awareness for operators in a simulated control room as part of the human-system interface design and evaluation process. Data regarding these two constructs will ultimately be submitted as evidence of design adequacy for certification by the Nuclear Regulatory Commission. The NRC calls for state-of-the-art assessment. At present, definitions of these constructs vary, as do the techniques for their assessment.

Based on an extensive review of related work in human factors, and with a special emphasis in the nuclear power domain, we believe state-of-the-art assessment includes the application of multiple measures. Subjective measures offer simple application and data analysis, and, due to their wide use, are the most likely to afford comparison with other systems. Physiological measures present practical challenges, but they promise greater objectivity and high time resolution. We believe that both categories should be represented in a state-of-the-art assessment. While mental workload and situation awareness are important considerations due to their impact on correct system operation, performance will also have to be measured in the simulator. Using performance to infer workload or situation awareness is problematic; we believe performance should be reported at the same level as, or even take precedence above, workload and situation awareness.

We recommended suites of measures at three budgetary levels, which we believed are generally desirable and which meet NuScale Power's particular program needs. We

then proceeded with the intermediate cost option, gaining first-hand experience with these measures in a simulated multi-modular reactor control room. As a result, we recommend NASA-TLX as a subjective measure of workload. With the assumption that state-of-the-art assessment of mental workload presently includes physiological techniques despite the challenges they pose, we recommend a wireless heart rate monitor and off-the-shelf analysis software as a low-cost approach. Although the pupil diameter measure of workload showed promise in our literature review, its sensitivity to a number of factors will make data collection in a simulated environment challenging. Even if good data is obtained, it may not be easy to explain the results confidently. This may be true for physiological measures in general. We did not fully follow the SACRI method of objectively measuring situation awareness, but our experience with the inventory identified a number of issues. Nevertheless, we believe such an objective freeze-based approach may represent the current state of the art. We recommend eye tracking for situation awareness assessment but concede that this should not be the sole method applied, because visual attention does not necessarily imply understanding.

The two constructs under consideration here, mental workload and situation awareness, should be generally useful in evaluating human-computer interactions, but they are most necessary for real-time, critical applications. We have proposed that multi-dimensional data from system-task-interface evaluations be presented together in a standard format for characterizing system safety and identifying problematic tasks, procedures, or user interface elements for re-design.

In proposing a multi-modular reactor control room, NuScale Power is venturing into new territory. Our experiment serves as an initial pilot study in simulator-based evaluation of this concept of operations. Although the numbers we obtained may not be applicable to an actual system, we have identified potential pitfalls, and have made suggestions regarding interface design for multi-module monitoring. Perhaps the most interesting result of our experiment from a human factors standpoint was that operators reported good situation awareness and moderate workload when following control procedures with highly automated system processing, but they neglected to monitor other reactors assigned to them. This implies design challenges related to encouraging high situation awareness across multiple independent units.

Given that we are unaware of any other existing empirical work in the human factors of multi-modular reactor control (see Plott et al., 2004), our literature review has identified the Uninhabited Aerial Vehicle (UAV) domain as a prime source for NuScale's operating experience review. UAV designers and evaluators are asking similar questions to NuScale Power, regarding the inter-related issues of levels of automation, operator-to-unit ratio, mental workload and situation awareness.

We have noted from the literature that high levels of automation present design challenges. The user interface design must find a balance to ensure that the operator stays "in the loop" without resulting in information overload. There is some evidence that this may be more easily achieved with intermediate levels of automation. Although adaptive automation is being applied successfully in some domains, we feel this design solution

introduces too much uncertainty and additional complexity to system design, evaluation, risk analyses and operation in the nuclear power domain.

There are several opportunities for further investigation. Due to the goals of our pilot study, we observed interesting trends, but we were unable to make statistically substantiated claims. A full study with an actual system design and more realistic tasks should address questions similar to ours. As noted by Hwang et al. (2008), such considerations as workload and situation awareness will need to be investigated over the duration of a normal shift, rather than for a few shorter, more engaging scenarios.

Due to various issues, our pupil diameter data was problematic. Based on a follow-up review of the literature, we believe pupil diameter should be investigated as a potential objective indicator of operator fatigue or disengagement in long-shift monitoring tasks. Also, based on one of our results, it appears worthwhile to compare pupil diameter and NASA-TLX in a carefully controlled HCI experiment with a range of task difficulty levels. Positive correlation would provide convergent validity of the two dissimilar measures, while an inverted-U relationship might suggest task disengagement at high workload levels (cf. Rowe et al., 1998), and thus potential evidence for a NASA-TLX-based redline for overload.

At present, there are no widely accepted redlines for mental workload or situation awareness. Few have even been proposed. The user interface design process should incorporate an understanding of these constructs, and their assessment should help to identify problematic task-interface combinations in the simulator. However, there is no fixed target, at which a human factors program can report that a system design is

objectively "adequate." High-level comparisons can be made with other systems, to the extent that empirical results are publicly shared. Task performance provides the most direct indication of system design adequacy, and is at present the most meaningful way of specifying redlines for acceptability.

A traditional reactor control room could be used as a benchmark for acceptability, as Ha et al. (2007) have suggested. For mental workload, intermediate levels appear to be optimal. This finding is in agreement with NUREG-1791: staffing plan validation analyses should "demonstrate that the staffing plan does not result in either excessively high or minimal workload demands on control personnel" (Persensky et al., 2005, p. II-10-8). However, it remains unclear whether NuScale Power's design efforts for typical steady-state operations should be more concerned with preventing underload, due to the low event rate, or overload, due to the demands of monitoring a complex system. That is, does the widely observed vigilance decrement in monitoring tasks spring from the fact that the operator must work hard to maintain attention or that he is underworked? This is a complex issue: Low event rates and low task load can result in boredom and decreased physiological arousal, suggesting underload (see, e.g., Pattyn et al., 2008). At the same time, the operator may need to exert mental effort to maintain attentional focus on the task and thus becomes frustrated, but mental effort and frustration are both contributors to heightened subjective mental workload. In this sense, underload and overload could be viewed as "two sides of the same coin" (Pattyn et al., 2008, p. 376) for such conditions. Our limited experiment could not address this question adequately, but the results appear to fit with this complex formulation. Participants reported being relatively engaged at

intermediate to somewhat high levels of workload, but the physiological measures seemed to reflect decreasing arousal over time, despite generally increasing levels of task load. Therefore, questions remain regarding the long-term, uneventful monitoring of a complex system. Perhaps the most important finding from our literature review and our pilot study is the need to apply measures from various categories (e.g., subjective and physiological), as these seem to reflect different aspects of workload under such conditions.

In light of widely known and carefully scrutinized past system incidents and accidents, regulatory agencies such as the United States Nuclear Regulatory Commission and the Federal Aviation Administration recognize the importance of considering human-system interactions in ensuring safe system operation. However, there is no universally agreed upon method for analyzing important aspects of operator interactions with a complex system. Altogether, this work represents a small, first, necessary step toward answering these questions for NuScale Power's Human Factors Engineering program. From the human-computer interaction standpoint, real-time, critical applications present interesting and challenging design problems. With human lives and well-being potentially at stake, the empirical evaluation of proposed interface designs merits great care and thought, as well.

BIBLIOGRAPHY

___. (2010). *FAA Aviation News, 49*.

Ahlers, M. M. (2009, October 27). *Pilots of wayward jet lose licenses*. Retrieved March 17, 2010, from http://www.cnn.com/2009/US/10/27/airliner.fly.by/index.html

Ahlstrom, U., & Friedman-Berg, F. J. (2006). Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics, 36*(7), 623-636.

Alexander, A. L., Nygren, T. E., & Vidulich, M. A. (2000). *Examining the relationship between mental workload and situation awareness in a simulated air combat task.* (AFRL-HE-WP-TR-2000-0094). Air Force Research Laboratory, Human Effectiveness Directorate, Wright-Patterson AFB, OH.

Associated Press. (2009, October 26). *NTSB: Wayward pilots were working on laptops*. Retrieved March 17, 2010, from http://www.msnbc.msn.com/id/33483228/

Baldwin, C. L. (2003). Neuroergonomics of mental workload: New insights from the convergence of brain and behaviour in ergonomics research. *Theoretical Issues in Ergonomics Science, 4*(1/2), 132.

Berka, C., Levendowski, D. J., Cvetinovic, M. M., Petrovic, M. M., Davis, G., Lumicao, M. N., et al. (2004). Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. *International Journal of Human-Computer Interaction, 17*(2), 151-170.

Berka, C., Levendowski, D. J., Ramsey, C. K., Davis, G., Lumicao, M. N., Stanney, K., et al. (2005). *Evaluation of an EEG workload model in an Aegis simulation environment.* Biomonitoring for Physiological and Cognitive Performance during Military Operations, Orlando, FL.

Blackwell, A. F., & Green, T. R. G. (1999). *Investment of attention as an analytic approach to cognitive dimensions.* Collected Papers of the 11th Annual Workshop of the Psychology of Programming Interest Group (PPIG-11).

Boles, D. B., & Adair, L. P. (2001). The Multiple Resources Questionnaire (MRQ). *Human Factors and Ergonomics Society Annual Meeting Proceedings, 45*, 1790-1794.

Braby, C. D., Harris, D., & Muir, H. C. (1993). A psychophysiological approach to the assessment of work underload. *Ergonomics, 36*(9), 1035-1042.

Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology, 45*(4), 602-607.

Brat, I. (2010, February 17). *The emotional quotient of soup shopping*. Retrieved February 17, 2010, from http://online.wsj.com/article SB10001424052748704804204575069562743700340.html?mod= WSJ_hpp_sections_business

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology, 42*(3), 361-377.

Chi, E. H., Pirolli, P., Chen, K., & Pitkow, J. (2001). Using information scent to model user information needs and actions and the Web, *Proceedings of the SIGCHI conference on Human factors in computing systems*. Seattle, Washington: ACM.

Colle, H. A., & Reid, G. B. (1998). Context effects in subjective mental workload ratings. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 40*(4), 591-600.

Colle, H. A., & Reid, G. B. (2005). Estimating a mental workload redline in a simulated air-to-ground combat mission. *International Journal of Aviation Psychology, 15*(4), 303-319.

Collier, S. G., & Folleso, K. (1995). SACRI: A measure of situation awareness for nuclear power plant control rooms. In D. J. Garland & M. R. Endsley (Eds.), *Experimental analysis and measurement of situation awareness* (pp. 115-122). Daytona Beach, FL: Embry-Riddle University Press.

Conditions of licenses, 10 C. F. R. pt. 50.54 (revised periodically).

Contents of applications; technical information, 10 C. F. R. pt. 50.34 (revised periodically).

Dixon, S. R., & Wickens, C. D. (2003). *Control of multiple-UAVs: A workload analysis.* Paper presented at the 12th International Symposium on Aviation Psychology, Dayton, OH.

Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., Nikolic, D., & Manning, C. A. (1999). Situation awareness as a predictor of performance in en route air traffic controllers. (DOT/FAA/AM-99/3). Office of Aviation Medicine, Federal Aviation Administration.

Endsley, M. R. (1988). *Situation awareness global assessment technique (SAGAT).* Paper presented at the Aerospace and Electronics Conference (NAECON), 1988.

Endsley, M. R. (1993). *Situation awareness and workload: Flip sides of the same coin.* Paper presented at the Seventh International Symposium on Aviation Psychology.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*, 65-84.

Endsley, M. R. (1996). Automation and situation awareness. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 163-181). Mahwah, NJ: Lawrence Erlbaum Associates.

Endsley, M. R. (2000). Theoretical underpinnings of situation awareness: A critical review. In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness analysis and measurement* (pp. 3-32). Mahwah, NJ: Lawrence Erlbaum Associates.

Endsley, M. R., Bolte, B., & Jones, D. G. (2003). SA demons: The enemies of situation awareness. In *Designing for situation awareness: An approach to user-centered design* (pp. 31-42). New York: Taylor & Francis.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*, 381-394.

Endsley, M. R., Selcon, S. J., Hardiman, T. D., & Croft, D. G. (1998). A comparative analysis of SAGAT and SART for evaluations of situation awareness. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 42*, 82-86.

Endsley, M. R., Sollenberger, R., & Stein, E. (2000). *Situation awareness: A comparison of measures.* Paper presented at the Human Performance, Situation Awareness and Automation: User Centered Design for the New Millennium Conference, Savannah, GA.

Finomore, V. S., Shaw, T. H., Warm, J. S., Matthews, G., Riley, M. A., Boles, D. B., et al. (2008). Measuring the workload of sustained attention: Further evaluation of the Multiple Resources Questionnaire. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 52*(18), 1209-1213.

Fournier, L. R., Wilson, G. F., & Swain, C. R. (1999). Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: Manipulations of task difficulty and training. *International Journal of Psychophysiology, 31*(2), 129-145.

Gaillard, A. W. (1993). Comparing the concepts of mental load and stress. *Ergonomics, 36*(9), 991-1005.

Galster, S. M., Knott, B. A., & Brown, R. D. (2006). Managing multiple UAVs: Are we asking the right questions? *Human Factors and Ergonomics Society Annual Meeting Proceedings, 50*, 545-549.

Gevins, A., & Smith, M. E. (2003). Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science, 4*(1/2), 113.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Grier, R., Wickens, C., Kaber, D., Strayer, D., Boehm-Davis, D., Trafton, J. G., et al. (2008). The red-line of workload: Theory, research, and design. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 52*, 1204-1208.

Guhe, M., Liao, W., Zhu, Z., Ji, Q., Gray, W. D., & Schoelles, M. J. (2005). Non-intrusive measurement of workload in real-time. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 49*, 1157-1161.

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.* Paper presented at the IEEE Sixth Conference on Human Factors and Power Plants, 'Global Perspectives of Human Factors in Power Generation.'

Hallbert, B. P., Sebok, A., & Morisseau, D. (2000). A study of control room staffing levels for advanced reactors. (NUREG/IA-0137). Office of Nuclear Regulatory Research, US Nuclear Regulatory Commission.

Hancock, P. A., & Szalma, J. L. (2003). Operator stress and display design. *Ergonomics in Design, 11,* 13-18.

Hart, S. G. (2006). NASA-Task Load Index (NASA-TLX); 20 Years Later. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 50*, 904-908.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 139-183). Amsterdam: North Holland.

Healey, J. A., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *Intelligent Transportation Systems, IEEE Transactions on, 6*(2), 156-166.

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner, A. C., Zaklad, 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*, 429-439.

Hogg, D. N., Folleso, K., Strand-Volden, F., & Torralba, B. (1995). Development of a situation awareness measure to evaluate advanced alarm systems in nuclear power plant control rooms. *Ergonomics, 38*(11), 2394-2413.

Hornof, A. J., & Halverson, T. (2002). Cleaning up systematic error in eye-tracking data by using required fixation locations. *Behavior Research Methods, Instruments, & Computers, 34*(3), 592-604.

Hornof, A. J., Zhang, Y., & Halverson, T. (2010). *Knowing where and when to look in a time-critical multimodal dual task.* Paper presented at the ACM CHI 2010: Conference on Human Factors in Computing Systems, Atlanta, GA.

Hwang, S.-L., Yau, Y.-J., Lin, Y.-T., Chen, J.-H., Huang, T.-H., Yenn, T.-C., et al. (2008). Predicting work performance in nuclear power plants. *Safety Science, 46*(7), 1115-1124.

IEEE Std 845-1999. (1999). IEEE guide for the evaluation of human-system performance in nuclear power generating stations. New York: IEEE.

IEEE Std 1023-2004. (2005). IEEE recommended practice for the application of human factors engineering to systems, equipment, and facilities of nuclear power generating stations and other nuclear facilities. New York: IEEE.

Iqbal, S. T., Adamczyk, P. D., Zheng, X. S., & Bailey, B. P. (2005). Towards an index of opportunity: Understanding changes in mental workload during task execution, *Proceedings of the SIGCHI conference on Human factors in computing systems*. Portland, Oregon: ACM.

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004). Task-evoked pupillary response to mental workload in human-computer interaction, *CHI '04 extended abstracts on Human factors in computing systems*. Vienna, Austria: ACM.

ISO 10075-2. (1996). Ergonomic principles related to mental workload - Part 2: Design principles. Geneva, Switzerland: International Organization for Standardization.

Jones, D. G., & Endsley, M. R. (2000). *Can real-time probes provide a valid measure of situation awareness?* Paper presented at the Human Performance, Situation Awareness and Automation: User Centered Design for the New Millennium Conference.

Jones, D. G., & Endsley, M. R. (2004). Use of real-time probes for measuring situation awareness. *The International Journal of Aviation Psychology, 14*(4), 343-367.

Jou, Y.-T., Yenn, T.-C., Lin, C. J., Yang, C.-W., & Chiang, C.-C. (2009). Evaluation of operators' mental workload of human-system interface automation in the advanced nuclear power plants. *Nuclear Engineering and Design, 239*(11), 2537-2542.

Kavanagh, J. (2009, October 30). *Airline pilots struggle to stay focused*. Retrieved October 30, 2009, from http://www.cnn.com/2009/TRAVEL/10/28/pilots.cockpit/index.html

Klingner, J., Kumar, R., & Hanrahan, P. (2008). Measuring the task-evoked pupillary response with a remote eye tracker, *Proceedings of the 2008 symposium on Eye tracking research & applications*. Savannah, Georgia: ACM.

Learn, S. (2010, March 7). *Oregon State professor wants to help power a nuclear renaissance*. Retrieved March 17, 2010, from http://www.oregonlive.com/environment/index.ssf/2010/03/oregon_state_professor_wants_t.html

Levin, S., France, D. J., Hemphill, R., Jones, I., Chen, K. Y., Rickard, D., et al. (2006). Tracking workload in the emergency department. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 48*(3), 526-539.

Lew, R., Dyre, B. P., Werner, S., Wotring, B., & Tran, T. (2008). *Exploring the potential of short-time fourier transforms for analyzing skin conductance and pupillometry in real-time applications.* Paper presented at the Human Factors and Ergonomics Society Annual Meeting, New York.

Lin, C. J., Yenn, T.-C., & Yang, C.-W. (2010). Automation design in advanced control rooms of the modernized nuclear power plants. *Safety Science, 48*(1), 63-71.

Lin, T., Omata, M., Hu, W., & Imamiya, A. (2005). Do physiological data relate to traditional usability indexes?, *Proceedings of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future*. Canberra, Australia: Computer-Human Interaction Special Interest Group (CHISIG) of Australia.

Liu, D., Wasson, R., & Vincenzi, D. A. (2009). Effects of system automation management strategies and multi-mission operator-to-vehicle ratio on operator performance in UAV systems. *J. Intell. Robotics Syst., 54*(5), 795-810.

Luximon, A., & Goonetilleke, R. S. (2001). Simplified subjective workload assessment technique. *Ergonomics, 44*, 229-243.

Maxion, R. A., & Reeder, R. W. (2005). Improving user-interface dependability through mitigation of human error. *International Journal of Human-Computer Studies, 63*(1-2), 25-50.

Midy, M.-A., Jensen, C., & Park, Y. (2007). The commentator information system: A usability evaluation of a real-time sport information service. *Proceedings of the international conference on Advances in computer entertainment technology*. Salzburg, Austria: ACM.

Morad, Y., Lemberg, H., Yofe, N., & Dagan, Y. (2000). Pupillography as an objective indicator of fatigue. *Current Eye Research, 21*(1), 535-542.

Mouloua, M., Gilson, R., Kring, J., & Hancock, P. (2001). Workload, situation awareness, and teaming issues for UAV/UCAV operations. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 45*, 162-165.

Nachreiner, F. (1995). Standards for ergonomics principles relating to the design of work systems and to mental workload. *Applied Ergonomics, 26*(4), 259-263.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces, *Proceedings of the SIGCHI conference on Human factors in computing systems: Empowering people*. Seattle, Washington: ACM.

Norman, D. A. (1983). Design rules based on analyses of human error. *Commun. ACM, 26*(4), 254-258.

Norros, L., & Nuutinen, M. (2005). Performance-based usability evaluation of a safety information and alarm system. *International Journal of Human-Computer Studies, 63*(3), 328-361.

NRC. (2009, August 11). *Backgrounder on the Three Mile Island Accident*. Retrieved May 18, 2010, from http://www.nrc.gov/reading-rm/doc-collections/fact-sheets/3mile-isle.html

Nunan, D., Donovan, G., Jakovljevic, D. G., Hodges, L. D., Sandercock, G. R. H., & Brodie, D. A. (2009). Validity and reliability of short-term heart-rate variability from the Polar S810. *Med. Sci. Sports Exerc., 41*(1), 243-250.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In *Handbook of perception and human performance, Vol. 2: Cognitive processes and performance.* (pp. 1-49): Oxford, England: John Wiley & Sons.

O'Hara, J. M., Brown, W. S., Hallbert, B., Skraning, G., Persensky, J. J., & Wachtel, J. (2000). The effects of alarm display, processing, and availability on crew performance. (NUREG/CR-6691). Office of Nuclear Regulatory Research, US Nuclear Regulatory Commission.

O'Hara, J. M., Brown, W. S., Lewis, P. M., & Persensky, J. J. (2002). Human-system interface design review guidelines. (NUREG-0700, rev. 2). Office of Nuclear Regulatory Research, US Nuclear Regulatory Commission.

O'Hara, J. M., Higgins, J. C., Persensky, J. J., Lewis, P. M., & Bongarra, J. P. (2004). Human factors engineering program review model. (NUREG-0711, rev. 2). Office of Nuclear Regulatory Research, US Nuclear Regulatory Commission.

Parasuraman, R., & Wilson, G. F. (2008). Putting the brain to work: Neuroergonomics past, present, and future. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*, 468-474.

Patrick, J., James, N., Ahmed, A., & Halliday, P. (2006). Observational assessment of situation awareness, team differences and training implications. *Ergonomics, 49*, 393-417.

Pattyn, N., Neyt, X., Henderickx, D., & Soetens, E. (2008). Psychophysiological investigation of vigilance decrement: Boredom or cognitive fatigue? *Physiology & Behavior, 93*(1-2), 369-378.

Persensky, J., Szabo, A., Plott, C., Engh, T., & Barnes, V. (2005). Guidance for assessing exemption requests from the nuclear power plant licensed operator staffing requirements specified in 10 CFR 50.54(m). (NUREG-1791). Office of Nuclear Regulatory Research, US Nuclear Regulatory Commission.

Pirolli, P., & Card, S. (1995). Information foraging in information access environments, *Proceedings of the SIGCHI conference on Human factors in computing systems*. Denver, Colorado: ACM Press/Addison-Wesley Publishing Co.

Plott, C., Engh, T., & Barnes, V. (2004). Technical basis for regulatory guidance for assessing exemption requests from the nuclear power plant licensed operator staffing requirements specified in 10 CFR 50.54(m). (NUREG/CR-6838). Office of Nuclear Regulatory Research, US Nuclear Regulatory Commission.

Proctor, R. W., & Van Zandt, T. (2008). *Human factors in simple and complex systems* (2 ed., pp. 229-259, 269-270, 343-345). New York: CRC Press.

Reid, G. B., & Colle, H. A. (1988). *Critical SWAT values for predicting operator overload.* Paper presented at the Human Factors and Ergonomics Society Annual Meeting.

Reid, G. B., & Nygren, T. E. (1988). The Subjective Workload Assessment Technique: A scaling procedure for measuring mental workload. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (pp. 185-218). Amsterdam: Elsevier.

Rowe, D. W., Sibert, J., & Irwin, D. (1998). Heart rate variability: Indicator of user state as an aid to human-computer interaction, *Proceedings of the SIGCHI conference on Human factors in computing systems*. Los Angeles, California, United States: ACM Press/Addison-Wesley Publishing Co.

Rubio, S., Diaz, E., Martin, J., Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and Workload Profile methods. *Applied Psychology An International Review, 53*(1), 61-86.

Rueb, J. D., Vidulich, M. A., & Hassoun, J. A. (1994). Use of workload redlines: A KC-135 crew-reduction application. *International Journal of Aviation Psychology, 4*(1), 47-64.

Ruff, H. A., Calhoun, G. L., Draper, M. H., Fontejon, J. V., & Guilfoos, B. J. (2004). *Exploring automation issues in supervisory control of multiple UAVs.* Paper presented at the Second Human Performance, Situation Awareness, and Automation Technology Conference (HPSAA II), Daytona Beach, FL.

Salmon, P. M., Stanton, N. A., Walker, G. H., Jenkins, D., Ladva, D., Rafferty, L., et al. (2009). Measuring situation awareness in complex systems: Comparison of measures study. *International Journal of Industrial Ergonomics, 39*(3), 490-500.

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

Sarter, N. B., & Woods, D. D. (1995). How in the world did we ever get into that mode? Mode error and awareness in supervisory control. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 37*, 5-19.

Seeing Machines. (2009). *faceLAB 5*. Retrieved December 23, 2009, from http://www.seeingmachines.com/product/facelab/

Sirevaag, E. J., Kramer, A. F., Wickens, C. D., Reisweber, M., Strayer, D. L., & Grenell, J. F. (1993). Assessment of pilot performance and mental workload in rotary wing aircraft. *Ergonomics, 36*(9), 1121-1140.

Smith-Jackson, T. L., & Klein, K. W. (2009). Open-plan offices: Task performance and mental workload. *Journal of Environmental Psychology, 29*(2), 279-289.

Smolensky, M. W. (1993). *Toward the physiological measurement of situation awareness: The case for eye movement measurements.* Paper presented at the Human Factors and Ergonomics Society 37th Annual Meeting, Seattle, WA.

Spence, R. (2007). *Information visualization: Design for interaction.* (2 ed.). New York: Prentice Hall.

Svensson, E., Angelborg-Thanderz, M., Sjoberg, L., & Olsson, S. (1997). Information complexity-mental workload and performance in combat aircraft. *Ergonomics, 40*(3), 362-380.

Taylor, R. M. (1990). Situational Awareness Rating Technique (SART): The development of a tool for aircrew systems design. In *Situational awareness in aerospace operations (AGARD-CP-478)* (pp. 3/1-3/17). Neuilly-Sur-Seine, France: NATO - Advisory Group for Aerospace Research and Development.

Tharion, E., Parthasarathy, S., & Neelakantan, N. (2009). Short-term heart rate variability measures in students during examinations. *National Medical Journal of India, 22*(9), 63-66.

Theureau, J. (2000). Nuclear reactor control room simulators: Human factors research and development. *Cognition, Technology & Work, 2*(2), 97-105.

Tran, T. Q., Boring, R. L., Dudenhoeffer, D. D., Hallbert, B. P., Keller, M. D., & Anderson, T. M. (2007a). *Advantages and disadvantages of physiological assessment for next generation control room design.* Paper presented at the Human Factors and Power Plants and HPRCT 13th Annual Meeting, 2007 IEEE 8th.

Tran, T. Q., Garcia, H., Boring, R. L., Joe, J. C., & Hallbert, B. P. (2007b). *Human factors issues for multi-modular reactor units.* Paper presented at the Human Factors and Power Plants and HPRCT 13th Annual Meeting, 2007 IEEE 8th.

Tremoulet, P., Craven, P., Regli, S., Wilcox, S., Barton, J., Stibler, K., et al. (2009). Workload-based assessment of a user interface design. In *Digital Human Modeling* (pp. 333-342).

Tsang, P. S., & Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics, 39*(3), 358-381.

Tungare, M., & Perez-Quinones, M. A. (2009). Mental workload in multi-device personal information management. *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*. Boston, MA: ACM.

Veltman, J. A., & Gaillard, A. W. K. (1996). Physiological indices of workload in a simulated flight task. *Biological Psychology, 42*(3), 323-342.

Vicente, K. J., & Rasmussen, J. (1992). Ecological interface design: Theoretical foundations. *Systems, Man and Cybernetics, IEEE Transactions on, 22*(4), 589-606.

Vidulich, M. A., & Tsang, P. S. (1987). Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 31*, 1057-1061.

Vidulich, M. A., & Wickens, C. D. (1986). Causes of dissociation between subjective workload measures and performance: Caveats for the use of subjective assessments. *Applied Ergonomics, 17*(4), 291-296.

Warm, J. S., Dember, W. N., & Hancock, P. A. (1996). Vigilance and workload in automated systems. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 183-200). Mahwah, NJ: Lawrence Erlbaum Associates.

Welter, K. B., Bajorek, S. M., Reyes, Jr., J., Woods, B., Groome, J., Hopson, J., et al. (2005). APEX-AP1000 confirmatory testing to support AP1000 design certification (non-proprietary). (NUREG-1826). Office of Nuclear Regulatory Research, US Nuclear Regulatory Commission.

Werner, E. (2010, May 20). *Obama seeking more nuclear energy loan guarantees*. Retrieved May 26, 2010, from http://news.yahoo.com/s/ap/20100520/ap_on_bi_ge/ us_obama_energy_3

Wickens, C. D. (2002a). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3*(2), 159-177.

Wickens, C. D. (2002b). Situation awareness and workload in aviation. *Current Directions in Psychological Science, 11*(4), 128-133.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*, 449-455.

Wientjes, C. J. E. (1992). Respiration in psychophysiology: Methods and applications. *Biological Psychology, 34*(2-3), 179-203.

Wierwille, W. W., & Casali, J. G. (1983). *A validated rating scale for global mental workload measurement applications.* Paper presented at the Human Factors Society 27th Annual Meeting, Norfolk, VA.

Wilke, P. K., Gmelch, W. H., & Lovrich, Jr., N. P. (1985). Stress and productivity: Evidence of the inverted U function. *Public Productivity Review, 9*(4), 342-356.

Wilson, G. F. (2000). Strategies for psychophysiological assessment of situation awareness. In M. R. Endsley & D. J. Garland (Eds.), *Situation awareness analysis and measurement* (pp. 175-188). Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, G. F. (2002). An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *International Journal of Aviation Psychology, 12*(1), 3-18.

Wilson, J. R., & Funk, K. (1998). *The effect of automation on the frequency of task prioritization errors on commercial aircraft flight decks: An ASRS incident report study.* Paper presented at the Second Workshop on Human Error, Safety, and System Development, Seattle, WA.

Yeh, Y.-Y., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 30*(1), 111-120.

Zhang, Y., & Luximon, A. (2005). Subjective mental workload measures. *Ergonomia IJE & HF, 27*(3), 199-206.

Zijlstra, F. R. H. (1993). *Efficiency in work behaviour: A design approach for modern tools.* Delft University of Technology, Delft, The Netherlands.

APPENDICES

Appendix A

Modified SACRI Questionnaire for Condition 3A

mSACRI 3A

ID#_____

Unit 1

1. For unit 1, in comparison with the normal status, how would you describe the *pressurizer pressure*?
    a. greater than normal
    b. normal
    c. less than normal

2. For unit 1, in comparison with the normal status, how would you describe the *average core temperature (T_{ave})*?
    a. greater than normal
    b. normal
    c. less than normal

3. For unit 1, in comparison with now, predict how *steam generator pressures* will develop over the next few minutes.
    a. Both will increase
    b. One will increase
    c. No change
    d. One will decrease
    e. Both will decrease
    f. One will increase, one will decrease

Unit 2

4. For unit 2, in comparison with now, predict how *pressurizer level* will develop over the next few minutes. Pressurizer level will:
    a. increase
    b. stay the same
    c. decrease

5. For unit 2, in comparison with the normal status, how would you describe the *hot leg temperatures*?
    a. Both are greater than normal
    b. One is greater than normal
    c. Normal
    d. One is less than normal
    e. Both are less than normal
    f. One is greater, one is less than normal

6. For unit 2, in comparison with now, predict how *pressurizer pressure* will develop over the next few minutes. Pressurizer pressure will:
    a. increase
    b. stay the same
    c. decrease

Unit 3

7. For unit 3, in comparison with the recent past, how have *steam generator feed flows* developed?
    a. Both have increased
    b. One has increased
    c. No change
    d. One has decreased
    e. Both have decreased
    f. One has increased, one has decreased

8. For unit 3, in comparison with the normal status, how would you describe the *turbine power demand (electrical output)*?
    a. greater than normal
    b. normal
    c. less than normal

9. For unit 3, in comparison with now, predict how *steam generator levels* will develop over the next few minutes.
    a. Both will increase
    b. One will increase
    c. No change
    d. One will decrease
    e. Both will decrease
    f. One will increase, one will decrease

10. For unit 3, in comparison with the recent past, how has *neutron flux* developed?
    a. increased
    b. stayed the same
    c. decreased

Unit 4

11. For unit 4, in comparison with the recent past, how has *core power* developed?
    a. increased
    b. stayed the same
    c. decreased

12. For unit 4, in comparison with the recent past, how have *cold leg temperatures* developed?
    a. Both have increased
    b. One has increased
    c. No change
    d. One has decreased
    e. Both have decreased

Summary

13. How would you describe the current status of each unit? For each unit, check all applicable boxes.

	Normal, steady-state operation	Undergoing planned, operator-initiated state change	Abnormal conditions / issues / unexpected fluctuations	Shutdown – not in operation	Emergency
Unit 1					
Unit 2					
Unit 3					
Unit 4					

14. How would you rate the current health of each unit? Put one check mark in each row.

	5 – Excellent	4 – Good	3 – Fair	2 - Poor	1 - Emergency
Unit 1					
Unit 2					
Unit 3					
Unit 4					

15. How would you rate your current awareness of the plant?

    a. 5 - Excellent
    b. 4 - Good
    c. 3 - Fair
    d. 2 - Poor
    e. 1 - Emergency

Appendix B

Script for Post-Experiment Interview (Semi-Structured)
ID #: _____

Experiment-Related Questions

1. Task Demands

    a. Rank the tasks you were asked to perform during the experiment by difficulty from 1 (most difficult) to 3 (least difficult). That is, each task should be assigned a unique number: {monitoring one unit, monitoring four units, procedural operation with four units}

        i. Why did you rate ____ the most difficult? What was challenging about it?

        ii. Why did you rate ____ the easiest?

        iii. Compare the workload and situation awareness of monitoring four vs. procedural operation with four.

    b. What parts of the experiment were the most/least stressful and why?

    c. Were any tasks difficult but not stressful, or any tasks stressful but not difficult? Why?

    d. Which operating rules were the hardest to follow and why?

    e. Did you try to use the scan pattern we recommended? If so, was it helpful?

    f. Did you use the plots? What percentage of the time? Were they useful?

g. How did your mood change from the beginning to the end of the experiment?

    i. E.g. more stressed or more comfortable, more alert or more tired, more engaged or more bored?

h. In terms of keeping your eyes/mind focused, how difficult was it to monitor four modules and why? (1- very easy to 10-very hard)

i. In terms of "keeping everything straight" (i.e., not confusing modules), how difficult was it to monitor four modules and why? (1 to 10)

j. How would you compare this concept of operation (i.e., your role as monitor of multiple highly-automated reactors) with your previous experience?

k. How well did the digital HSI support you in your tasks? What was positive/negative about the control and display interface?

2. Physiological Measures

    a. Did the heart rate band get in the way of your assigned tasks? If so, which tasks, and how?

    b. Was the heart rate band ever annoying or uncomfortable?

    c. How distracting was it to know that your eye activity was being tracked? (1-10)

    d. Was the eye tracker calibration procedure annoying? Why?

    e. How distracting was it to know that you were being videotaped? (1-10)

f. [As applicable, ask about observed divergence between subjective and physiological measures. For example…] Typically, we would expect that when workload increases, your heart rate would increase and your NASA-TLX numbers would increase. However, in [observed task], this was not the case. Tell me about the workload/stress/pressure/demands you experienced in that task. What may have caused your heart rate to [increase/decrease], even though you felt like workload was [low/high] relative to the other conditions?

3. Survey Instruments

a. Did the disruptions (freezing the simulator) get in the way of your tasks? (e.g., did it ever cause you to lose situation awareness or your place in a procedure)

b. In what cases, if any, were the disruptions especially an issue (e.g., during monitoring, shortly after an alarm, during procedure-based operation)?

c. How annoying were the disruptions?

d. How well did the SA questions fit with your knowledge and mental model of the system? (i.e., did we ask things that you would typically monitor? Did we omit important parameters?)(arbitrary questions, 1-10, questions fit very well)

e. How well did the SA questions capture your awareness of the system state? (1-10)

f. How well did NASA-TLX capture your level of stress, workload, effort? (1-10)

g. Which dimensions contributed to your workload the most? Why?

h. Were the elements of your stress or workload that were not reflected in NASA-TLX?

Personal Information / Background

1. How old are you?
2. Do you have normal (20-20) vision without glasses or contacts?
3. Did you wear glasses or contacts during the session?
4. How much, if any, caffeine have you consumed today?
5. How much, if any, nicotine have you consumed today?
6. How many hours of sleep did you get last night?
7. Did you engage in strenuous physical activity before the session today?
8. Which of the following best describes your level of physical fitness?
    a. Poor – exercise rarely or never, Fair – exercise occasionally, Good – exercise regularly, Excellent – serious athlete in training
9. Height and weight (for heart rate monitoring equipment)?
10. How many years of training and/or experience do you have in PWR operation? Describe.
11. Have you used a reactor simulator before?
12. Have you used PCTRAN, the simulator used in this study, before today?
13. How familiar was the simulator?

14. What is your level of education in Nuclear Engineering?

Opportunity for Debriefing

1. Do you have any other feedback about your experience, or suggestions for similar studies in the future? (instructions, tasks, setup, procedures, questions, etc.?)

2. Finally, do you have any questions about the experiment that we can address?