The Development and Evaluation of Attention and Situation Awareness Measures in Nuclear Process Control Using the Rancor Microworld Environment

A Dissertation

Presented in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

with a

Major in Experimental Psychology

in the

College of Graduate Studies

University of Idaho

by

Thomas A. Ulrich

Major Professor: Steffen Werner, Ph.D.

Committee Members: Ronald Boring, Ph.D., Rajal Cohen, Ph.D.; Brian Dyre, Ph.D.

Department Administrator: Todd Thorsteinson, Ph.D.

December 2017

Authorization to Submit

This dissertation of Thomas A. Ulrich, submitted for the degree of Doctor of Philosophy with a Major in Experimental Psychology and titled "The Development and Evaluation of Attention and Situation Awareness Measures in Nuclear Process Control Using the Rancor Microworld Environment," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: ____________________________________ Date: _____________

Steffen Werner, Ph.D.

Committee Members: ____________________________________ Date: _____________

Brian Dyre, Ph.D.

____________________________________ Date: _____________

Rajal Cohen, Ph.D.

____________________________________ Date: _____________

Ronald Boring, Ph.D.

Department

Administrator: ____________________________________ Date: _____________

Todd Thorsteinson, Ph.D.

Abstract

Nuclear process control is a complex domain requiring human factors research to ensure the safe, reliable, and efficient operation of nuclear power plants. The aging fleet of U.S. nuclear power plants are undergoing control room modernization. This control room modernization effort focuses on upgrading the human system interfaces with advanced digital interfaces. To aid this effort, the Rancor Microworld was developed as a platform to examine human factors and related psychological constructs in the complex nuclear process control domain. Rancor was designed for use with novice student populations to augment full-scope simulation studies using expert operator participants. Students and operators were evaluated while they completed the Rancor Microworld process control task. Several aspects of human performance were examined, including primary process control task performance, situation awareness, and patterns of attention. Situation awareness, a key element of complex process control, was assessed with SAGAT-like and SACRI-like freeze probes. Patterns of attention were assessed using a traditional eye tracking fixation measure. Additionally, a new simple attention-acknowledgement measure was proposed, developed, and evaluated to provide researchers with another method of assessing attention within the complex process control domain. Results from this work suggest students and operators demonstrate similar performance with a noticeable decrement for operators, likely due to negative expertise transfer, during the first trial. Operators demonstrated better situation awareness than students, which suggests there is at least one key difference between the novice student and expert operator populations. The eye tracking fixation-based measure and the attention-acknowledgement measure of attention both assessed similar patterns of attention during the Rancor Microworld task. Additionally, both measures were sensitive to different patterns of attention for the various process control subtasks, which indicates the attention-acknowledgement measure is an effective method for assessing patterns of attention in a complex nuclear process control domain.

Acknowledgements

This work would not have been possible without the financial support of the Nuclear Science and Technology program at Idaho National Laboratory. I am especially grateful to Dr. Ronald L. Boring who has been supportive of my academic work and provided valuable mentoring as well as many opportunities to gain invaluable human factors and nuclear industry experience including working with fellow researchers and industry collaborators, interacting with operators, and visiting nuclear facilities.

I would also like to express my deepest thanks to Dr. Steffen Werner who inspired and mentored me throughout my doctoral journey. Without his invaluable mentoring and guidance, I never would have attempted this monumental endeavor that spanned five years of my life. Furthermore, the meticulous attention to detail maintained the necessary stringent scientific practices necessary to enable this project to succeed. I am forever and deeply indebted.

Roger Lew played a central role in this project, in particular in the development of the Rancor Microworld. His substantial assistance and the knowledge he shared were crucial to the development of the simulation. Indeed, substantial portions of the simulator were built based on code he developed for previous simulation efforts.

Aubrey Milatz was fundamental to the laboratory organization and data collection process. Her diligence and competence were crucial to maintaining records and data. Her sarcastic sense of humor was tremendously helpful to maintain laboratory morale throughout this long series of experiments. Lastly, I would like to thank Marshall Masingale and Cody Anderson for their excellent work collecting and scoring data.

Dedication

To Kira, my loving wife, best friend, and coconspirator. Also, to my parents, Mary and Tom, who provided the original inspiration for all my scientific endeavors.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Roughly 20% of the electricity produced in the U.S. is generated by nuclear power plants (NPPs). The current fleet of NPPs are undergoing digital control upgrade projects as part of their efforts to extend the lifespan of the plants beyond their original 40-year licensing periods (Boring et al., 2012a; 2012b; 2013). Many of the plants rely on analog technology that is quickly becoming obsolete and costly to maintain. As a part of this digitization, control room modernization focuses on integrating these digital control systems and interfaces into existing control rooms through a series of scheduled upgrades. Central to these upgrades is ensuring operators maintain high-levels of situation awareness (SA) to support increased safety, reliability, and efficiency while operating the NPPs (O'Hara et al., 2011). A significant amount of human factors research, focusing specifically on NPP control rooms must be performed to ensure the safety and usability of the new digital systems.

Human factors practitioners must work with plant managers, engineers, operations personnel, and control room operators to identify the specifications for the new digital interfaces. Designs can be created based on these specifications, but these designs must be thoroughly tested and evaluated to ensure they adhere to stringent safety and reliability requirements. In some respects, guidance is lacking when it comes to the design, implementation, and evaluation of digital human-machine interfaces (HMIs). Nuclear control human factors research came to a stall after nuclear power fell out of favor in the 1980s and 1990s following several high-profile incidents, such as the Three Mile Island Accident (Kemeny, 1979). Control room modernization efforts have spurred the need for a significant amount of human factors research, which must be performed to ensure the safety and usability of the new digital systems within the control room. Indeed, the topic of control room modernization has undergone a significant increase in publications as evidenced by a literature search conducted for the term “control room modernization,” yielding 84 results for the 2000 to 2010 time period and 181 results for the 2010 to the present time period, which represents a doubling of research efforts on the topic.

NPP operators use a complex HMI in the form of a control room with control boards containing thousands of indicators and controls (Boring et al., 2012a; 2012b; 2013). Operators face the challenging task of monitoring and controlling the plant to ensure safe, efficient, and reliable electrical power production. The operators’ process control task places considerable demands on the operators due to the complex relationships between the multitudes of systems involved with the nuclear power production process. As such, the United States Nuclear Regulatory Commission (U.S. NRC) issues regulatory documents to provide utilities with strategies to ensure good human factors are followed when implementing any new interface changes to the main control room. This guidance is provided in the form of NUREG-0711, “Human Factors Engineering Program Review Model,” which outlines the process of analyzing human factors needs and addressing those needs in the interface to ensure good usability for the operators controlling the plant (O’Hara et al., 2011). NUREG-0711 identifies many ways to address the human factors issues during the design and evaluation phases of the upgraded interface, but in particular, NUREG-0711 explicitly states the importance of assessing SA in the current system and ensuring that the new system maintains or exceeds the current level of SA.

Of the numerous approaches to evaluating HMI interactions, SA is one of the most prominent methods employed (Endsley, 1995a). Furthermore, SA is typically measured during interface evaluations as evidenced by NUREG-0711's specific mandate to evaluate and ensure it is maintained during the design process (O'Hara et al., 2011). Acquiring SA requires many perceptual and cognitive constructs, such as attention, visual perception, working memory, and decision-making. All of these underlying concepts play a role in building SA, but attention is particularly relevant, since it drives the selection of important information from the plethora of status and control information displayed across the control boards (Wickens et al., 2008). Furthermore, patterns of attention aid in measuring SA since only attended-to-information can be incorporated into the operator's SA, and therefore, the pattern of attention informs what information comprises the operator's SA. Due to attention's prominent role in acquiring SA, a new measure of SA based on an attention-acknowledgement measure is the focus of this research effort. The following section will describe SA as a construct and how it is employed as a methodology for research in complex domains.

Chapter 2. Situation Awareness

SA is a widely acknowledged but not uncontroversial concept within human factors (Burns et al., 2008). In general, SA refers to the knowledge of what is going on around you. The field of aviation coined the term during the First World War (Patrick et al., 2006). Heightened SA was proposed as a causal factor for a small number of pilots shooting down a disproportionately large number of enemy aircraft (Hartmann & Secrist, 1991; Kelly et al., 1979). Since then, researchers have applied SA to a variety of domains involving dynamic complex tasks at both the individual- and team-level. SA is researched in a number of diverse domains, including military command and control (Durlach, Kring, & Bowens, 2008; Mathews & Beal, 2002), process control (Burns et al., 2008; Kaber & Endsley, 1998), driving (Gugerty, 1997; Walker, Stanton, & Young, 2008), and anesthesiology (Gaba et al., 1995). There are multiple competing models with different conceptualizations of the SA construct. SA has been used to refer to knowledge in working memory (Bell & Lyon, 2000), the knowledge of a situation (Endsley, 1995a), externally directed consciousness (Smith & Handcock, 1995), and cognitive activities required in dynamic, event-driven, and complex tasks (Sarter & Woods, 1995). Most conceptualizations of SA include an aspect of attention, but some even define SA within the context of attention as exemplified by Fracker's conceptualization, "the knowledge that results when attention is allocated to a zone of interest at a level of abstraction" (1991, p. 102). Of the various competing models and conceptualizations, Endsley's model is the most widely accepted within the human factors field and a number of practitioners have adopted her three-level SA model (1995a, 1995b). Unlike other models, Endlsey's explicitly differentiates between the process of acquiring knowledge and the knowledge itself. According to Endsley's model *situation awareness* is the knowledge itself, while *situation assessment* is the process of acquiring that knowledge (Endlsey, 1995a). Henceforth, SA refers to situation awareness or knowledge of the situation itself.

Endsley's three-level model defines SA 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" (1995a, p. 37). Level I consists of an awareness of individual elements via perception of those elements in the environment, Level II consists of comprehending the situation by synthesizing the individual elements into a meaningful composite representation, and Level III consists of projecting the future states of the integrated elements representation. The model is hierarchically organized such that each level requires knowledge from the level below it, while each higher level represents a greater understanding of the situation. This hierarchical organization provides the model with the capability to isolate breakdowns in SA that occur in a particular situation or system, as well as to evaluate how the system's interface supports each level of SA.

Assessing Situation Awareness

SA has received considerable attention within applied research conducted in industrial settings. Numerous techniques to measure SA in these applied industrial settings have been developed. A review of SA measures by Salmon et al. (2009) classified 30 different specific measures into four main categories: self-ratings, observer ratings, freeze probe technique, and real-time probe technique. Each of the 30 specific measures consists of some variant of these four main techniques developed for a specific domain or to evaluate SA within the framework of a specific model.

Situation Awareness Methods

The various SA measures can be categorized into freeze probe, real-time probe, self-rating, and observer rating techniques. The following sections describe each of these techniques and include an example use case.

Self-rating is the most basic technique used for evaluating SA. In the self-rating technique, individuals rate themselves on multiple dimensions of their subjective SA after completing a task. The Situation Awareness Rating Technique (SART) is an example of a self-rating technique that uses ten dimensions to quantify SA (Taylor, 1990). Individuals rate themselves on a seven-point Likert scale for each SA dimension. The self-rating technique is easy to administer and does not impose any primary task intrusion since it is completed after the task has commenced. The SART has a simplified version to provide a more rapidly administered technique using only three dimensions, which is termed the 3D SART (Taylor, 1990). Yet another prominent self-rating technique is the Situation Awareness Rating Scales Technique (SARS), which is similar to the SART but developed specifically for military aviation (Waag & Houck 1994).

Observer ratings are another widely accepted subjective measure used for SA evaluation. Subject matter experts (SMEs) simply observe and rate participants' SA while the participant performs the task (Salmon et al., 2009). The SMEs code participants' SA on predefined observable behaviors in an attempt to increase the rating consistency. Observer ratings are advantageous since they require minimal primary task intrusion and can be conducted in real-life industry settings without interference. The Situation Awareness Behavioral Rating Scale (SABARS) is an example of an observer rating technique. Matthews and Beal (2002) used SABARS to measure the SA of infantry soldiers in field-training exercises.

The freeze probe technique is the most widely used objective SA measurement tool. The freeze probe technique consists of administering SA-related queries while the task is temporarily suspended (Endsley, 1995b). The individual's responses reported during the freeze probe are compared to the actual system state at that point in time, as defined by the experimenter, to yield an overall SA score for a task. The primary benefit of the freeze probe technique is the immediate, objective SA assessment at different time-points throughout the simulation, as opposed to measurements of SA at the end of a trial. The Situation Awareness Global Assessment Technique (SAGAT) is an example of a well-known freeze probe technique designed with queries that specifically evaluate SA at each of the three levels of Endsley's SA model (Endsley, 1995b). Queries from the SAGAT developed for use in aviation consist of questions concerning a pilot's knowledge of the aircraft's airspeed, altitude, attitude, and location. There is more evidence correlating performance with the SAGAT freeze probe technique than any other SA measurement (Salmon et al., 2009).

The real-time probe technique is another SA measure that relies on providing participants with SA-related queries (Salmon et al., 2009). Unlike the freeze probe technique, the real-time probe does not suspend the simulation. This technique was developed to mitigate primary task intrusion due to suspending the simulation. The content of the answers and the response time in providing the answers are used to generate a score for the level of SA. The Situation Present Assessment Method (SPAM) is an example of a real-time probe technique used to evaluate SA in air-traffic controllers (Durso et al., 1998). The SPAM is remotely administered over the telephone to air-traffic controllers. The response times for correct answers are used to assess the level of SA. A shorter response time reflects the air-traffic controller with a high-level of SA, since the air-traffic controller can mentally recall the information or efficiently direct his or her attention towards the necessary indicator to retrieve the information quickly.

Current Situation Awareness Methodological Disadvantages

Though each of the methods mentioned in the previous section have use cases that make them advantageous, each has disadvantages as well. The self-rating task collects participants' subjective ratings after the trial has ended, which forces participants to rely on memory to report his or her recalled SA. Therefore, human memory limitations may prevent participants from accurately reporting their SA. Self-ratings may be distorted due to an individual's biased perception of their performance during the simulation (Endsley, 1995b). Furthermore, individuals must remember multiple mental states when rating themselves on the various dimensions, which confounds self-ratings with long-term working memory and recall abilities. When rating the different dimensions of SA, individuals must condense dynamic moments of SA throughout the task to a single average value for each SA dimension. Collapsing the dynamic moments into a single average value can result in substantial bias due to memory distortions (Kahneman, 1973). The most troubling disadvantage for subjective self-report ratings are that they may not necessarily correlate with performance during the task. Participants can and often do rate themselves highly across the SA dimensions; however, their objectively measured performance may have been, in fact, poor. Similarly to the self-report ratings of SA, observer ratings also suffer from subjective bias. Bias can affect both the observation in terms of what behaviors are noted, as well as how those behaviors are rated, which are the two primary disadvantages of the observer rating technique. Furthermore, it may be difficult for the observer to disambiguate between SA and the actions taken such that the rating simply reflects the individual's performance instead of the underlying SA. Inter-rater reliability may pose an issue, which makes replication of experiments challenging without the original SME involvement. Thus, making comparisons between studies and disciplines is quite difficult.

The more objective freeze probe and real-time probe techniques avoid the subjective issues associated with the self-report and observer rating techniques. However, the most significant disadvantage of the probe technique stems from primary task intrusion. In the case of the freeze probe technique, the task is suspended, which disrupts the flow of the task and can destroy the realism for simulated tasks. The real-time probe also suffers from primary task intrusion because responses must be provided concurrently while completing the main task. Due to these various disadvantages, new methods for assessing SA are worth pursuing, though another issue concerning SA must be addressed as well. Any investigation into SA inherently requires a scenario in which an individual performs a task that can be examined. The next section will describe the different environments employed to support scenarios and tasks necessary for SA investigations.

Situation Awareness and Nuclear Process Control

SA is particularly important for nuclear process control because the complexity of the systems represents a significant challenge for understanding the state of the plant. Indeed, even at the lowest level of Endsley's model, in which the operator must identify the relevant pieces of information, the challenge to the operator is considerable due to multiple control boards with more than a thousand indicators and controls.

Loss of SA can be associated with numerous incidents in the nuclear industry. In particular, the Three Mile Island accident was the biggest nuclear-related accident in U.S. history. In this accident (Kemeny, 1979), a series of automatic plant actions caused a pilot relief valve to open to reduce pressure building up in the primary reactor system. The valve was supposed to close once this pressure had been reduced; however, the valve remained stuck open. The indication in the control room displayed the valve as closed since its binary state was linked to the command signal sent to close the valve. Though the signal was sent to close the valve, the valve position itself had no indication, and therefore, its position was inferred from the command signal. The opened valve allowed coolant to rapidly escape the core in what is referred to as a loss of coolant accident. The operators used the pressurizer to infer the level of the coolant in the core; however, with the opened valve this assumption was faulty and the core had insufficient coolant. The cladding on the fuel ruptured and the plant suffered a core meltdown.

Due to accidents like the one at Three Mile Island, a number of studies have been conducted on how teams of operators maintain SA while operating a plant. Typically, nuclear control room studies consist of anywhere from one to six operating crews completing multiple scenarios (Carvalho, Vidal, & de Carvalho, 2007; Ulrich, Boring, & Lew, 2014). At the onset of each scenario, the operators are provided with a crew brief detailing the current state of the plant, such as startup, steady-state, or cooldown, along with a handful of key parameter values. A fault is then injected to observe how the crew identifies the problem and returns the plant to a safe state. Common scenarios include steam generator tube ruptures, loss of coolant, loss of onsite power, and rapid downpower (Ulrich, Boring, & Lew, 2014).

A number of different SA methods have been employed in nuclear process control, but the most common consist of observer and self-ratings. For example, Carvalho, Vidal, & de Carvalho observed crews during shift changeover to capture the verbal content of the relieved crew briefing the starting crew (2007). Their results suggest that this debrief is critical for crews to build their initial SA of the plant prior to beginning their monitoring and control tasks during their respective shifts. A self-rating-based technique was developed specifically for nuclear process control by Hogg et al. (1995) based on the SAGAT. The Situation Awareness Control Room Inventory (SACRI) entails querying the operator multiple times throughout the scenario with specific questions targeted towards known important process parameters as evident in the following example, “In comparison with recent past, how have the temperatures in the hot legs of the primary circuit developed?” (Hogg et al., 1995, p. 2411). Recent past refers to the past three minutes of the scenario and the response options are general, such as increase, decrease, or remain the same. Each query took no longer than five minutes depending on how long the operator took to ponder the question, and these lengths of time were found to not interrupt the flow of the scenario. The SACRI demonstrated sensitivity to differentiate between scenario events, such as the onset of a fault when SA typically suffers, the individual operators, and interface design manipulations.

SA is used in complex domains as described in the prior sections, and is an important concept for understanding how individuals in complex environments understand the system they are interacting with in the context of the environment to achieve good performance. However, there are also many other cognitive phenomena that are also relevant to complex environments and human performance. Notably, attention is a cognitive construct that is critical to complex process control and also related to SA. The subsequent section discusses the construct of attention and how it relates to complex process control. Furthermore, the way attention is researched is also described.

Chapter 3. Attention

Attention and Situation Awareness in the Nuclear Operators' Tasks

The main control room is a complex HMI comprised thousands of indicators and controls displayed on large metal control boards spanning the entire outer perimeter of the main control room. The control room allows the operators to coordinate the multitude of distributed systems and personnel spanning across the plant from a centralized location. Operators face the challenging task of supervisory monitoring and controlling plant systems to ensure safe, efficient, and reliable electrical power production. The operators' process control task places considerable demands on the operators due to the complex relationships between the multitudes of systems and their associated complex interactions involved with the nuclear power production process. Researchers have used several different approaches to evaluate the HMI in nuclear process control in their efforts to design better control rooms and more efficient operations. Of these different approaches, SA is one of the most prominent methods employed (Endsley, 1995b). SA is a broad-reaching concept, which has garnered significant criticisms due to its somewhat nebulous and all-encompassing conceptualizations (Flach, 1995). Numerous definitions of SA exist, but at its most fundamental level, SA simply refers to "…knowing what is going on around you," (Endsley, 2000). SA subsumes many perceptual and cognitive constructs, such as visual perception, working memory, decision-making, and attention. All these underlying psychological concepts play a role in the acquisition and maintenance of SA, but attention serves as a particularly important contributing factor to support SA acquisition within nuclear control rooms. In particular, visual attention allows operators to select regions of indicators and controls and seek out particular pieces of information from the enormity of information displayed across the control boards. Auditory attention is important since there are auditory alarms in the control room; however, there are only a handful of auditory alarms and each is sufficiently salient that they are readily detectable. An operators' visual attention is under considerably more demand than his/her auditory attention, which is why this paper focuses primarily on the visual domain of attention.

Vigilance, selective, and divided forms of attention are all involved to varying degrees as operators monitor and control the NPP from the main control room. The operators' task involves a degree of vigilance since there are some low-activity periods in which the operators must scan the boards in anticipation of an event during normal operations. However, many interesting research questions pertain to how attention plays a role in an operators' tasks during abnormal plant situations in which the operator must assess faults to determine which system or systems faulted, determine the root cause of the fault, diagnose and verify the symptoms of the fault, manipulate the controls to return the plant to a normal operating conditions, and then monitor the plant to ensure the fault was effectively corrected and that the systems have returned to steady-state normal operation. This series of tasks engages a visual search to identify pertinent indicators and controls and selective attention capabilities to filter out the irrelevant indicators and controls as operators seek to isolate the relevant information. Divided attention is required to shift attention between the multiple tasks the operator must perform simultaneously. Vigilance plays a diminished role in relation to these other forms of attention, and therefore, will not be discussed further. Attention is important for more than simply SA and nuclear process control as evident by the volume of research conducted on the topic spanning back more than a century.

The Psychological Concept of Attention

Attention is a fundamental component shaping our conscious experience as we interact with the world around us. In *The Principles of Psychology*, William James provides an intuitive definition that captures the essence of attention (1890, p. 404):

> Everyone knows what attention is. It is the taking possession by the mind in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought...It implies withdrawl [*sic*] from some things to deal effectively with others, and is a condition which has a real opposite in the confused, dazed, scatterbrained state.

The concept of attention is largely intuitive for most people since we are able to consciously direct it in basic everyday endeavors, such as walking down the street or reading a book. Attention in its most mundane and basic form is referred to as vigilance attention. Vigilance attention entails focusing on a task while waiting for an event to occur or a stimulus to appear over a prolonged period of time. For example, security personnel perform carry-on baggage screening with an x-ray machine to detect dangerous items passengers might bring onto an airplane. Sensitivity to accurately detect the stimulus diminishes over time, especially for rare occurring stimuli. Pioneering vigilance work conducted by Mackworth demonstrated, in possibly the most boring experiment imaginable, that individuals suffer a vigilance decrement over time (1948). Reserve Air Force cadet participants were instructed to watch the second hand of a clock over a two-hour time period. Participants were asked to identify and report the second-hand taking a double-step, which occurred between approximately one- to five-minute intervals during the two-hour watch period. At the end of the first half-hour, participants failed to report 15% of the second-hand double-steps. During the last half-hour of the two-hour watch period, the decrement had magnified with participants failing to report approximately 30% of the second-hand double-steps. Fatigue and resulting decreased arousal are primary factors driving the vigilance decrement (Fisk & Schneider, 1981), but the decrement is also due to habituation resulting from repeated exposure to the background events of the vigilance task during time periods in which the stimulus is not present (Mackworth, 1968). The response bias can be explained by signal detection theory in terms of a shift in the participants' criteria towards not reporting a double-step since this response results in more correct responses due to the low event rate (Nevin, 1969).

Vigilance attention in its pure form is not overly interesting within the context of nuclear process control because operators are commonly performing several tasks that require them to move throughout the control room and attend to many different visual displays. As such, vigilance will not be referred to again in lieu of discussing more relevant forms of attention. More complex forms of attention involve consciously and unconsciously directing our focus towards important stimuli throughout the environment. Our senses are bombarded by a bewildering enormity of stimuli that we do not perceive since attention filters out the vast majority of the unneeded stimuli we encounter every day. There is simply far too much information in even a simple environment for the human mind, despite its remarkability, to perceive and process in any meaningful way without attention.

The research community broadly defines attention in terms of the functions it serves in various situations as it drives our conscious experience of the world. Each of these general functions has several accompanying theories and methodologies used to examine psychological mechanisms of attention. The three broad categorizations of attention include selective attention, divided attention, and visual search. Selective attention, divided attention, and visual search all play a role in the operators' process control task; however, in relation to other important cognitive concepts such as SA, some play larger roles than others. In particular, selective attention plays a crucial role in how operators allocate attention.

Selective Attention

In the majority of real-world situations, including the operators' process control task, the visual system is assailed with a constant stream of visual stimuli competing for our attention. Selective attention refers to the ability to focus on a particular stimulus within the environment while ignoring other competing stimuli. Selective attention allows us to shift our attention and hold it on a particular stimulus of interest. Three primary classes of selective attention models include the filtering models and pre-attentive and attentive processing models.

The filtering model is one of the oldest and most influenced models of selective attention, which emerged from the auditory work performed by Broadbent (1958) and later by Driver (2001). In auditory shadowing experiments, participants listened to two auditory message streams of information and were required to attend to one while ignoring the other (Cherry, 1953; Broadbent, 1958; Moray; 1959). To ensure that participants were in fact attending to one auditory message, they were instructed to shadow the message they heard, which consisted of rapidly repeating the content of the message. Broadbent (1958) found that in order for efficient selective attention, the two messages must be physically distinct, such as originating from a different location or spoken with a different pitch. More importantly, participants were unaware of much of the semantic content of the unattended message. For example, participants failed to report that the same word was repeated many times or the language of the unattended stream switched (Broadbent, 1958; Moray, 1959). These auditory shadowing experiments led Broadbent to proposed the filter model of selective attention in which only attended information is passed on to more advanced stages of cognitive processing (1958). Broadbent's model predicted the unattended message would be filtered out before it entered conscious awareness; however, experimentally there were key elements of the unattended auditory message that appeared to pass through the filter, such as the participants' names or a word paired with a prior electrical shock (Moray, 1959), were noticeable. In light of this evidence, Treisman modified the original Broadbent (1958) filter model to account for these properties that passed through the filter (1960, 1969). Treisman's model states that the stimuli are attenuated rather than completely filtered out, which allows some processing to occur at a deeper level for stimuli with particular properties, though this is more an exception than a rule since most stimuli are filtered (1960, 1969). Though these early filtering models are based on auditory experiments, they are precursors to visual filtering models.

Broadbent's and Treisman's filtering models operate within the auditory domain. There are similar filtering models that operate within the visual domain. The saliency model attributed to Itti et al. (1998) describes a visual filtering mechanism that is analogous to those in Broadbent's and Treisman's auditory-filtering mechanisms. According to the saliency model, selection of visual stimuli is driven by a bottom-up processing of physical characteristics of the stimuli such as luminance, color, location, and orientation (Itti, Koch, & Niebur, 1998). A saliency map topographically encodes stimuli in terms of their similarity to surrounding stimuli within the visual scene. A stimulus with a high-conspicuity in relation to surrounding stimuli will result in a higher activation of the topographical feature map, which in turn draws attention towards the location of the stimulus. The less activated stimuli are then inhibited, which further enhances the attentional capture of the conspicuous stimulus. This initial attentional capture effect based on the conspicuity of a stimulus in relation to the other stimuli present in the visual scene gives rise to the concept of multiple stages of processing associated with visual attention.

The pre-attentive and attentive processing models (Van der Heijden, 1996) of selective attention map onto the filtering models since the first pre-attentive stage based on relative stimuli conspicuity consists of parallel processing, which is sensitive to certain gross physical attributes, while the attentive stage of processing deals with more advanced features and their integration (Treisman & Gelade, 1980). Pre-attentive and attentive models of attention are directly related to visual search. Visual search is one of the key categories of attention, but since it is also a research paradigm in its own right, a later section will discuss visual search models, such as the pre-attentive and attentive class of models. Additionally, a large body of research has been conducted on selective attention using the flanker interference paradigm. A detailed discussion on how selective attention mechanisms allow us to filter out some stimuli, while focusing on another, will be discussed in their empirical context in a later section on the flanker interference paradigm.

Divided Attention

While performing process control tasks, operators engage extensively in concurrent tasks. Even while performing a single procedure, operators must monitor and control the process, while at the same time following the procedure to ensure that each of the detailed steps is performed correctly. When multiple procedures are being completed, which is more often the case than not, the operators must further divide their attention between these various tasks. This imposes a significant challenge on the operators' perceptual and cognitive abilities, in which attentional capabilities can become overwhelmed and result in operator error. Outside of the operators' task, the general population also engages in multiple tasks extensively, and therefore, divided attention is a prevalent attention function we use throughout our lives.

Divided attention, within visual attention research, refers to the selection of multiple stimuli or regions within the visual field. Divided attention can be characterized within either spatial or temporal dimensions. Within the spatial dimension, divided attention refers to tracking multiple stimuli at different spatial location points within the visual field simultaneously. Shaw & Shaw found evidence for simultaneously attending to multiple locations separated by either space or an object simultaneously (1977), though the vast majority of evidence suggests that attention is only directed to a single region at a given point in time (Posner, 1980).

Within the temporal dimension, divided attention refers to the shifting of attention serially between two or more stimuli through time. Many studies have been performed to examine divided attention in simple and complex situations, such as laboratory-based stimulus-response experiments and talking on a cell phone while driving (Pashler, 1994; Strayer & Johnson, 2001; Strayer, Drews, & Johnson, 2003). The dual-task interference paradigm's defining characteristic is the marked performance decrement resulting from two, even very simple, tasks performed simultaneously versus in isolation (Pashler, 1994). The decrement is attributed to a phenomenon known as the psychological refractory period in which the response to a second stimulus or second task is delayed since the first stimulus or task is still undergoing mental processing (Telford, 1931). The stimulus onset asynchrony of the two stimuli or tasks defines the magnitude of the dual-task interference with shorter stimulus onset asynchronies yielding greater interference and longer delays in response to the second stimulus.

Resource models of attention were proposed to explain dual-task interference. The resource models view attention as a limited supply of processing capacity that can be distributed to various tasks (Kahneman, 1973; Wickens, 1991). Within the visual domain, there are specific resource models that differ in how the pool of resources can be divided between locations within the visual field. For example, the model proposed by Shaw & Shaw (1977) permits the pool of resources to be differentially distributed to multiple locations simultaneously, due to empirical evidence that suggested participants could distribute attention to various locations based on the expected probability for a target to appear within that location. Other resource models make different assumptions, such as the resource pool being allocated in either a focused or distributed nature depending upon the stage of processing (Jonides, 1981). Though each of these models make different assumptions, the important component of these models is the concept that a finite pool or processing capacity is available at a given time and must be drawn upon by multiple tasks if they are performed simultaneously. Kahneman proposed the original resource theory model, which posits that all tasks compete for the same pool of processing capacity (1973). Wickens refined Kahneman’s model by incorporating multiple resource pools associated with the different stimulus and response modalities as outlined in the human information processing theory (Isreal et al., 1980). For example, a visual stimulus with a physical response, such as a button press, will generate less interference on an auditory stimulus and verbal response task, since the two tasks are occurring within separate stimulus and response modalities.

In addition to accounting for dual-task interference, resource models of attention also reiterate the importance of selective attention in governing the attentional shifts required to perform two or more tasks concurrently. Due to the limited pool of resources, selective attention is important to direct those resources in the most efficient manner. Selective attention helps direct the operators' attention to select visual regions required for a task, which is then re-engaged to direct the dived attention between concurrent tasks.

Useful Conceptual Metaphors of Attention

Divided, visual search, and selective attention all entail moving the locus of attention throughout the visual scene. This moving capability of attention raises the following questions surrounding mechanisms that would allow attention to operate as a moveable region of selection within the visual field (Cave & Bichot, 1999):

- What is the size and shape of the region of selection?
- Is the size of the selection region fixed or can it be mentally adjusted?
- Is the region of selection contiguous or can it select disparate regions?
- How quickly can the region of selection be shifted to another region within the visual field?
- What processing, if any, occurs outside the region of selection?

To address these issues, the spotlight and zoom lens analogies were proposed. These analogies provide an intuitive framework for conceptualizing the mechanisms that allow "attention to operator" as a selection mechanism within the visual field.

Spotlight Metaphor

Visual attention is operationally defined as the selection of a physical area or object within the visual scene for enhanced perceptual and cognitive processing (Cave & Bichot, 1999). This definition of attention is incomplete because it fails to define mechanisms for how areas or objects undergo selection, and the nature of this selection region in terms of size, shape, contiguity, and temporal characteristics. To address the shortcomings of this definition, a spotlight metaphor of attention was postulated to operationally define constructs that afford testable hypotheses. Though the idea of a spotlight metaphor has existed for centuries, Posner is credited with coining the term. The spotlight is defined as an area that "enhances efficiency of detection of events within its beam" uniformly, while everything outside the beam is not selected and does not receive any benefits of enhanced efficiency (Posner, Snyder, & Davidson, 1980, p. 172). This definition is the strict interpretation of the spotlight metaphor, which is also referred to as the "obligatory spotlight" because it defines the region of selection with a concrete boundary. Furthermore, the obligatory spotlight conceptualization asserts that all stimuli falling within the spotlight beam undergo equal enhanced-processing, while everything outside the beam undergoes no enhanced-processing. The basis for this assertion stems from experimental findings in which participants demonstrated faster reaction times to detect or identify a target presented at a pre-cued location (Eriksen & Hoffman, 1974). The faster reaction times following the pre-cue indicate support for the spotlight analogy since it establishes that there are boundaries surrounding the beam of attention. Furthermore, the cue serves to orient the attention spotlight at the location in which the stimulus will be presented. This eliminates any time required to move the spotlight of attention to the stimulus that would otherwise be required without a cue. The cue serving as a mechanism to orient attention raises another key element of the spotlight metaphor, which deals with shifting the spotlight of attention.

Another core component of the spotlight metaphor is the ability to move the spotlight throughout the visual field. The spotlight can be directed through the visual field to scan for interesting or highly salient stimuli. Two complimentary mechanisms support the shifting of attention through the visual field. The first is termed overt attention or overt-orienting of attention and considers the spatial locus of attention as yoked to the visual focal point (Rizzolatti et al., 1987). Overt-orienting of attention can be subdivided into voluntary and reflexive forms (Kastner, 2014). Voluntary orienting of attention is also referred to as endogenous or sustained attention since the attention-allocation is governed by internal mechanisms within the individual. Conversely, reflexive orienting attention is referred to as transient or exogenous attention because the attention-allocation is captured by an external stimulus within the visual field of the individual. Eye movements in the form of saccades towards the intended locus of attention are associated with both transient and sustained attention (Salvucci & Goldberg, 2000). The second general form of attention-orienting is referred to as covert attention and is differentiated from overt attention in that covert attention is not yoked to the visual focal point of the individual (Kowler, 2011). Instead, overt attention allows the individual to attend to locations away from the focal point, such as regions near the periphery of the visual field. Furthermore, covert-orienting is not associated with eye movements, and therefore, the locus of attention is not considered strictly yoked to gaze position (Posner, 1980). Covert- and overt-orienting of attention function as complimentary systems to direct attention throughout the visual field. Covert-orienting primarily serves the role of detecting interesting or salient stimuli throughout the visual field, while overt-orienting maintains attention on a particular location for additional focused-processing. Covert attention is closely linked to the motor control of saccades in which the covert-orienting detects the stimulus, while the ocular motor response moves the eye to fixate on the stimulus (Salvucci & Goldberg, 2000). In the process control setting, covert-orienting is not as important as overt because the operators are trained to scan the control boards to detect pertinent information using overt-orienting.

The spotlight analogy is valuable for conceptualizing attention in many situations; however, there are some mechanisms of attention that operate outside of the analogy. For example, inattentional blindness is a phenomenon in which a visual stimulus goes unnoticed despite it residing within the fixation area typically yoke to the spotlight of attention. A stimulus can go unnoticed for two reasons. First, the locus of attention can be uncoupled from the current fixation point during covert attention (Carrasco & McElree, 2001). Furthermore, even if the unnoticed stimulus falls within the visual fixation, it is sometimes not noticed. For example, the famous inattentional experiment involving a gorilla walking amongst a group of people passing a basketball between each other (Simons & Chabris, 1999). Participants are instructed to count the number of passes, which causes the participants to focus their attention on the basketball object within the visual scene. The gorilla even moves in close proximity to the basketball, which brings the gorilla within the spotlight beam of attention. However, since the gorilla is of little interest due to the task demands of counting the number of passes, selective attention filters it out of conscious perception. A similar phenomenon, known as change blindness, operates in a similar manner. Changes made to a visual scene can go unnoticed if attention is never directed to the object that undergoes the change in much the same way as inattentional blindness results in the failure to detect a stimulus (Simons & Levin, 1997). The spotlight views attention from a spatial perspective, but a significant amount of evidence suggests attention can operate in other non-spatial dimensions. Object-based attention explains the inattentional blindness phenomenon in this gorilla experiment, since the participants are focusing their attention on the basketball object, even when the gorilla is co-located in the same area within the visual field.

The spotlight analogy also does not account well for effects of the spotlight beam size or how it can be adjusted to accommodate stimuli of varying size (Cave & Bichot, 1999). For example, if a small stimulus is expected, the size of the beam could be restricted to focus solely on the target stimuli and eliminate any interference from nearby. Conversely, the size of the spotlight would need to be enlarged to accommodate larger objects. Indeed, object-based selection is an entire separate area of research that has received considerable attention. Attention has the capacity to select for objects as evidenced in experiments in which object cues enhance processing and lead to shorter reaction times for a target presented within that object, as opposed to a target presented outside the object (Harms &

Bundesen, 1983). The spotlight analogy assumes that attention is distributed evenly throughout the selected region and applies equally regardless of the size of the spotlight region. This assumption doesn’t always correspond with existing data, since experiments cueing participants to expect a target of a given size, followed by a target of a different size, demonstrate reaction times increase in proportion to the size of the expected target (Larsen & Bundesen, 1978; Cave & Kosslyn, 1989). As the spotlight increases in size, the beneficial enhanced-processing appears to diminish as a tradeoff for a larger area of selection. The zoom lens analogy of attention addresses this issue of the attention resolution in relation to the selection region size.

Zoom Lens Analogy

The spotlight metaphor makes the assumption that attentional resources are restricted to the beam of attention aimed at a specific selected region. All smaller locations or objects within this beam receive equal attentional resources. By analogy, the acuity differences present in the eye, specifically the high-acuity in the foveal region and the decreasing acuity as eccentricity increases towards the periphery, provide some basic analogous biological evidence that attention circuitry is also probably not uniform (Cave & Bichot, 1999). Returning to attention, it is difficult to reconcile the spotlight metaphor in situations with highly salient stimuli that pop-out for immediate involuntary detection by individuals (Connor, Egeth, & Yantis, 2004). If attentional resources are purely restricted to the spotlight, then the pop-out effect would not be possible. Some attentional resources must be distributed throughout the visual field in addition to the likely majority of attention reserved for enhanced-processing within the spotlight beam.

The zoom lens analogy was proposed to address these shortcomings from the spotlight metaphor’s assumption of a uniform distribution of attention within the beam. It was also proposed to overcome the obvious disadvantage of the spotlight metaphor in which it cannot account very well for attentional capture outside of the beam. The zoom lens subsumes the spotlight analogy conceptualization and extends it by adding a resolution component. The zoom lens adds an additional distributed mode of operation in which attentional resources are dispersed throughout the visual field with varied resolutions (Eriksen & Yeh, 1985). Within the zoom lens model, the distributed and focused modes of operation of attention are viewed as a continuum (LaBerge, 1983). On one end of the continuum, a small focused spotlight beam of attention is able to select a small, one-degree of visual angle or less region for a large boost in processing efficiency, as evidenced by shorter reaction times for detection or identification of targets. On the other end of the spectrum, attention can be distributed over a much larger region of the visual field; however, during this form of operation, the boost in processing efficiency is much less pronounced. Therefore, a tradeoff exists between the size of the selection region and the magnitude of the enhanced-processing associated with attentional resources.

Whether the spotlight or zoom lens analogy is used to describe the manner in which an operator directs his or her attention while monitoring and controlling the plant, the limited capabilities of attention create a challenging situation for the operator. The operator must direct his or her attention to critical pieces of information in a control room containing many thousands of pieces of information. This is a daunting task and it is worth a moment to marvel at the ability for operators to function in such an environment so successfully with the impeccable safety record that U.S. NPPs maintain. Operators must be strategic in their attention-allocation to acquire the critical information at the correct time. The next section on eye tracking will explain measuring operator attention-allocation in the nuclear control room.

Attentional Correlates of Situation Awareness

Within the framework of Endsley's three-level SA model, attention is vital at all levels of the model, but it is particularly important for Level I, which consists of the perception of relevant pieces of information within the environment (1995a). The actual patterns of attention exhibited by an individual as he or she interacts with the display are measured through time, and that pattern is compared against an estimated optimal pattern of attention. The optimal pattern of attention is estimated based on the characteristics of the interface and the task demands. For example, a source of information should be sampled in a specific sequence or at a specific time in relation to other display elements for the participant to accurately assess the system state and build their SA appropriately. Models relating attention to SA, such as Wickens' Attention-Situation Awareness (A-SA) model and specifically the SEEV subcomponent of the model, can be employed to translate attention-allocation to reflect an individual's SA within Levels I and II of Endsley's three-level model (Wickens et al., 2008).

Wickens' A-SA model includes a component he refers to as the SEEV model, which serves to map attention to SA (Wickens & McCarley, 2008). The SEEV model is an applied selective attention model of how an operator directs his or her attention during visual attention tasks in complex work environments (Wickens et al., 2003). The SEEV model joins together elements from visual search models, such as feature integration (Treisman & Gelade, 1980) and guided search (Wolfe, 1994). It also incorporates attentional capture models of saliency (Itti, Koch, & Niebur, 1998). Lastly it includes both bottom-up and top-down processing that drives visual attention (Itti, 2000). Collectively, these elements come together to form the four components of the model, which form the acronym SEEV, as shown in Table 3.1.

Table 3.1 SEEV model parameters used to determine amounts of attention that would optimally be allocated to areas of interest with the visual scene of a task.

SEEV Model of Attention	
Saliency	Bottom-up driven attention capture effects or "pop-out"
Effort	Effort to shift attention to an area of interest
Expectancy	Information bandwidth within a location or area of interest
Value	Utility of the information based on its **R**elevance and **P**riority

The SEEV Model

For each source of information within the display, the probability of the operator attending to that source is a combination of saliency, effort, expectancy, and value. The saliency of each source is rated based on how intrinsically noticeable the item is in terms of its physical characteristics, such as size, shape, color, and location. The effort required to focus attention on a source is based on the centrality of the source within the display in addition to the other task demands. The effort to attend to a source is a negative factor that reduces the probability of attention being allocated towards that source. As task demands and competition for attention resources increases, the inhibitory effect of effort also increases. The centrality is not simply defined by the physical position within the display, but also accounts for the positioning of attention in the display during the task. For example, if a source is near a cluster of other information sources that are used extensively during the task, it will require less effort to shift attention towards it since it is near the locus of attention throughout the task. Operationally, the model identifies scan paths, which consist of transitions between sources of information and their respective distances which are used to quantify the effort involved with shifting attention to each source. Expectancy refers to the rate at which new information is conveyed by the source and any contextual cueing that would alert the operator of new information present at the source. There is little need to shift attention towards the source at a higher frequency than it updates, since no new information is gained from doing so. The value of the item is based on the utility it provides to accomplish the task goals. The value can be defined in numerous ways, such as the cost of missing new information at a source, the benefit of integrating the information at a source with other information in the display, and in some cases, the source is simply important in relation to the task.

SEEV Model Relationship to Situation Awareness

Wickens & McCarley (2008) refers to the SEEV model as static since each of the parameter values do not change once they are assigned to a display element. This can yield a matrix of probabilities of time spent attending to each screen element, referred to as percent dwell-times, over the entire task or series of tasks the display is used to complete. From these probabilities, the static version of the SEEV model can be used to map attention distribution for an individual over the duration of the task (Wickens & McCarley, 2008). A dynamic version of the SEEV model can be achieved by incorporating time-windows and recalculating the parameters based on the system state within each subsequent time-window. The dynamic version is much more complicated, since it also requires analyzing the scan paths to determine the transitions between AOIs. An additional workload module can also be incorporated into this dynamic version to account for changes in workload, which can exhibit influence on attention and specifically the effort parameter that represents the effort required to switch attention from one display element to another with the magnitude of this effort increasing in step with the distance between the two display elements. The dynamic provides more fine-grained analysis than the static, but at the cost of significant additional complexity in the analysis.

The predicted attention distributions account for how attention should be allocated towards display elements based on each elements' relative saliency, effort, expectancy, and value. Therefore, the SEEV model becomes a useful tool to understand how attention should be allocated throughout the task based purely on the system and interface design. This prescriptive model of how attention should be allocated is representative of an individual exhibiting high-SA since the matrix of distributions represents an optimal attention distribution in which the individual attends to the key elements with percentage dwell-times that afford maximum information acquisition. The attention distributions generated by the SEEV model are equivalent to the information contained within Levels I and II of Endsley's SA model, which concerns the perception of elements within the environment and understanding of those elements in relation to each other (1995a).

Attention-Situation Awareness Model Empirical Evidence

Wickens' A-SA model has been validated in a few attention studies involving aviation and driving tasks. In one study, Wickens et al. applied the attention module SEEV model to visual fixation data of pilots flying simulated landings aided by different display interfaces under two different levels of visibility (Wickens et al., 2004; McCarley et al., 2004). The model predictions of calculated attention distributions were compared against the actual pilots' fixation distributions. The experiment also included an objective measure of Level I SA measured using probe questions concerning traffic awareness during the landing approach. The pilots were probed to determine if they detected two off-normal events comprised of a rogue blimp requiring a flight path deviation and runway alignment depicted in their advisory support system necessitating the pilots to adjust their approach to land on the actual runway. Additionally, the different display interfaces were compared against each other to determine which afforded the best performance; however, the details of that comparison are omitted since they are of little relevance to the actual validity of the SEEV model to evaluate attention-allocation. The results from these comparisons demonstrated strong correlations between the SEEV model's predicted attention distributions and the observed fixation distributions of the actual pilots. Fixation distributions were also compared against tracking error, which is representative of general flight performance on the landing task. An optimality score was calculated, which consisted of the correlation between the predicted SEEV model and the actual pilots' fixation distributions. Pilots with higher optimality scores also exhibited better mean path tracking than pilots with lower optimality scores, indicating that performance is related to attention patterns and the associated construct of Level I and Level II SA. The objective measure of SA assessed with the probe questions differentiated between the experimental conditions concerning the interfaces and were not compared against either the predicted or observed pilot attention patterns. It is unclear how the probe questions were administered to the pilots, though it appears likely they were administered during the debrief after each experimental trial. The probe technique used by Wickens et al. is similar to other freeze probe techniques used to formally evaluate SA; however, the probes are typically administered during the experimental trial in a freeze probe technique (Salmon, et al., 2009). The following section will describe the traditional eye tracking method for measuring attention-allocation.

The SEEV Model as a Qualitative Model

Though the SEEV model was intended to be used as a method to quantitatively model optimal attention patterns, the model itself has undergone little validation in regard to formulating appropriate and psychologically valid weights for each of the parameters when used to calculate the optimal attention pattern. Indeed, it is unclear as to how the experiments performed by Wickens et al. specifically assigned the weights for each of the parameters and they themselves reported that it took several attempts to fit the model to the data (Wickens et al., 2004; McCarley et al., 2004). For the purposes of measuring attention within process control, it is outside the scope of this research effort to validate the model. However, the model's variables of saliency, effort, expectancy, and value are all valid psychological constructs that should govern how attention is directed; therefore, the SEEV mode provides a useful framework for accounting for differences in SA as a result of differences in patterns of attention. Henceforth, all references to the SEEV model assume the model purely as a qualitative model as a means to describe the patterns of attention in relation to task demands and interface characteristics.

Measures of Attention

Reaction Time

Reaction times are commonly used as a measure of attention. As described in a previous section on the spotlight metaphor of attention, when attention is cued to a location the reaction time to detect a stimulus at that location is shorter than when attention is not cued to that location (Eriksen & Hoffman, 1974). The cuing of attention to the location eliminates the need to shift attention and the time required to do so, which has the net effect of reducing the overall reaction time for the response to detect or identify stimuli. Within the context of complex process control, the reaction time would be reduced when attention is already directed to pertinent information. For example, an operator will demonstrate shorter reaction times to diagnose a system fault if his or her attention is directed to the location of the interface conveying the fault than when attention is directed elsewhere. The time-course for reaction times in traditional measures of attention is quite small in the magnitude of hundreds of milliseconds, but within the realm of complex process control, the reaction times are more on the magnitude of seconds and minutes.

Accuracy

Within the context of attention, accuracy refers to the ability to correctly identify and report a visual or auditory object presented to a participant. Typical attention research paradigms employing the accuracy measure present stimuli for brief durations to cued and non-cued locations. Participants demonstrate better accuracy when the stimuli was briefly displayed in a cued as opposed to non-cued or invalidly cued locations (Eriksen & Eriksen, 1974; Jonides, 1981). This improved accuracy is associated with channel enhancement, which refers to attention enhanced information gathering or signal-enhancement within the locus of attention (Prinzmetal, McCool, & Park, 2005). Through enhanced-processing, the perceptual representation of the stimuli tends to be more veridical, and therefore, participants can more accurately report the identity of the stimuli.

Eye tracking

Eye tracking is a useful technique to capture where an individual's eyes fixate at a given point in time (Hauland, 2003), as well as the pattern of eye movements associated with information processing across displays during human-computer interactions. Researchers have employed eye tracking techniques to measure attention-allocation in complex systems across numerous domains including aviation (Sarter, Mumaw, & Wickens, 2007), surface transportation (Ji & Yang, 2002), and medicine (Law et al., 2004). Eye tracking has also been specifically used in a number of studies concerning human-computer interaction and nuclear process control (Kovesdi et al., 2015). The research authored by Kovesdi et al. (2015) is particularly relevant because it addresses the specifics of conducting eye tracking research within a full-scope simulation of a NPP main control room. To achieve eye tracking in this complex three-dimensional environment, the authors relied on wearable eye tracking units in the form of glasses.

Eye tracking measures attention and its allocation through a visual scene based on the assumption that attention is typically yoked to the gaze position of the eyes (Duchowski, 2007). Eye tracking entails measuring the gaze position using infrared camera systems. In the most common technique employed in available commercial eye-trackers, the pupil and corneal reflection are used to calculate where the eye is pointed (Holmqvist et al., 2011). Head position is incorporated with the calculated direction of the eye to pinpoint a gaze location. Eye-trackers are capable of sampling at a wide range of speeds between 25 Hz and 2000 Hz (Holmqvist et al., 2011). A distinction is made between high- and low-speed sampling systems at 250 Hz. More accurate systems above this threshold are used for measuring smaller eye movements, such as microsaccads, tremors, and drift (Holmqvist et al., 2011). Most studies do not require this level of precision, and therefore, can make use of slower sampling eye tracking systems. The current line of research focusing on visual attention uses a Tobii X2-60 compact unit mounted under a desktop monitor and capable of sampling at 60 Hz, which has the necessary precision, one visual degree (Tobii, 2016), to adequately measure attention-allocation based on fixations and saccades (Salvucci & Goldberg, 2000; Goldberg & Wichansky, 2003).

A number of metrics can be used to characterize attention based on fixations and saccades. The most common metric for describing patters of attention is fixations per area of interest (Poole, Ball, & Phillips, 2005). Fixations per area of interest are useful for describing the pattern of attention across a physical space, such as an interface. Fixations per area of interest are also useful for determining the amount of attention dedicated to a given area of interest in relation to the rest of the visual scene, which is often accomplished by calculating the proportion of fixations within each are of interest within the visual scene. Fixation frequency is analogous to fixations per area of interest, but instead quantifies the pattern of attention over time (Kovesdi, Barton, & Rice, 2012). Fixation frequency is useful for describing patterns of attention during particular activities within the context of a larger task. Fixation durations are a metric used to describe the length of time in which an area is fixated. Fixation durations are useful for characterizing the level of effort required to process information within a particular area of interest under the assumption that longer durations are associated with increased cognitive effort (Marquart, Cabrall, & de Winter, 2015).

Eye tracking Disadvantages*.* Eye-trackers are a useful tool; however, they also suffer from several technical issues that make it challenging to use effectively in some environments. First, eye tracking suffers from numerous sources of errors that can lead to difficulty in accurately and reliably measuring each participants' gaze position. For example, a large portion of commercially available eye-trackers rely on infrared cameras to detect the pupil and corneal reflection of each eye to determine gaze position (Holmqvist et al., 2011). The process of capturing pupil and corneal reflection suffers when the camera cannot accurately capture either of these two items. Some individual differences that can interfere with this process include drooping eyelids that occlude the pupil, contact lenses that diffuse the corneal reflection, and mascara or eye makeup that generate false corneal reflections (Holmqvist et al., 2011). Additionally, for stationary camera-based systems, the head position must also be tracked along with the eye position, which suffers from other sources of errors, such as excessive participant movement and improper positioning away from the eye-tracker (Holmqvist et al., 2011). Both the eye and head position-tracking also suffer from interference based on lighting conditions (Holmqvist, et al., 2011). Beyond accurately recording the gaze data, the analysis can prove cumbersome for eye tracking. The data generated by eye tracking must undergo extensive processing to manipulate it into a more human digestible format necessary to answer research questions (Holmqvist et al., 2011). In addition to these general challenges associated with eye tracking, some environments pose specific challenges for eye tracking, such as the Human Systems Simulation Laboratory (HSSL), which is a full-scale, full-scope, reconfigurable, glasstop simulation environment capable of digitally representing nuclear control rooms for human factors design and usability evaluations (Boring et al., 2012a; 2012b; 2013). The HSSL supports touch inputs, which is important for representing the analog components across a digital environment to more naturally capture the realistic aspects of HMI between the operator and the control room. The HSSL platform has been primarily used at Idaho National Laboratory (INL) to perform applied research in collaboration with nuclear power utilities. As such, the timeline for running the experiments is tight and the cost of these experiments can be large (Ulrich, Werner, & Boring, 2015). With the brief time-course, it is important to collect the needed data as quickly as possible. Often the simpler subjective response measures provided by the operator participants provide the most valuable insights to improve upon the usability of new interface designs undergoing evaluation within the simulator (Ulrich, Werner, & Boring, 2015). The HSSL presents a challenge for eye tracking methodologies (Kovesdi et al., 2015), due to its complex three-dimensional environment containing many depth planes and spanning across 45 large displays with thousands of indicators and controls. Furthermore, several technical issues were encountered when using eye tracking in this environment, including battery life constraints for the portable eye tracking glasses and their processing units worn by the operator participants. Frequent recalibrations were necessary to perform between experimental trials to ensure the accuracy of the eye tracking. Furthermore, some eye tracking units use conflicting infrared camera systems and markers placed on the participant to determine head position. This type of infrared camera system for tracking head-position was discovered to be incompatible with the simulator touchscreen technology, which also relies on an infrared camera system embedded within bezels mounted over the displays to detect touch-positions. The touch-capabilities were rendered functionless when this eye tracking system was operating due to interference from conflicting infrared camera systems. From a human perspective of managing participants, the operators do not enjoy wearing bulky glasses-based systems that are compatible with the HSSL. These issues and others are encountered in other labs as well (Holmqvist et al., 2011), which provides the impetus to develop new measures that can answer the same questions in another fashion.

Discount Usability Approach for Nuclear Control Room Modifications

Conducting human factors research in complex environments is inherently challenging. First, the researcher must possess domain-specific expertise *and* knowledge of the system or access to experts with knowledge of the system to effectively design and conduct meaningful experiments. Second, it is challenging to carry out the actual experimental sessions, since the complexity entails a complicated dynamic system, all of which must be simulated, measured, and recorded to capture relevant human performance. Central to the recreation of the complex system human-computer interaction is this complicated process of simulating the complex system. Furthermore, the users of complex systems must be recruited and take time to participate in the experiment, which can be particularly challenging since these individuals are highly skilled and expensive personnel with full-time commitments to their actual work. NPP main control rooms represent particularly challenging complex sociotechnical systems for conducting human factors research, as illustrated in the following sections.

Access to highly restricted NPP main control rooms is not feasible for research purposes. The control rooms themselves are actively engaged in the control and monitoring of a plant. Furthermore, they adhere to stringent security protocols that preclude many researchers from stepping foot within these coveted spaces. Each power plant is also equipped with at least one full-scope training simulator that is an exact replica of the actual control room. This could be a suitable environment for some research, but these simulator facilities are extensively engaged in training and establishing qualifications for operators as mandated by Regulatory Guide 1.149 (U.S. NRC, 2001). Practically, the simulators are not well-suited to support research efforts because the metal control boards cannot easily be modified to accommodate testing different designs without permanent and license violating alterations necessary to mount new displays and move various components. As part of the U.S. NRC regulations, the simulator must be maintained as a functionally identical replica of the actual main control room, and therefore, these modifications are not possible. Even temporary modifications are infeasible because of the extensive use of the simulators for training and qualifications. An additional challenge is interfacing the robust but archaic analog control systems with modern computer simulations needed to develop prototypes for testing new interface designs.

Full scope simulators, such as the HSSL located at INL are dedicated facilities used almost exclusively for research efforts on control room modernization (Boring et al., 2012a; 2012b; 2013). Full-scope simulators are an invaluable tool for conducting human factors-based control room modernization research as demonstrated by numerous studies performed with full-scope simulators (Boring et al., 2015; Burns et al., 2008; Demas, Lau, & Elks, 2015). Full-scope simulators, such as the HSSL, are capable of digitally representing the main control room. As such, modifying the digitally represented control boards with prototypes of new interfaces becomes feasible. Furthermore, interfacing prototypes with the simulated plant model is made much simpler because the plant model contains digital representations of the analog components with application program interfaces (APIs) to support communication with the prototype interface software (Lew, Boring, & Ulrich, 2014).

Even with access to a full-scope simulator, it is still challenging to recruit operators to participate in the experiments. Operators are primarily responsible for keeping the plant up and running. Since plants run 24-hours a day, 7-days a week, the alternating crews of operators in charge of the plant are busily engaged. When operators are not operating the actual plant, they are also performing extensive training required to maintain their licenses. Thus, the operators have little time to devote for activities outside of their regular duties, such as participating in research experiments. Given these challenging circumstances for conducting research, it is important to make the best use of operators and simulator-time by performing experiments in the most efficient manner possible. The concept of discount usability offers an approach in which simple and easy to administer measurement techniques are used in lieu of complicated data rich techniques to make determinations about user performance.

Nielsen coined the term “discount usability” with an article titled, “Usability Engineering at a Discount” (Nielsen, 1989). The concept of discount usability refers to budget- and time-conscious usability evaluation techniques to arrive at high-quality user interfaces. The three main concepts encompassed by discount usability are simplified user testing, narrowed-down prototypes, and heuristic evaluation (Nielsen, 1989). Nielsen was referring to simple qualitative measures with his simplified user testing concept, which has proven to be a valuable approach in the nuclear process control domain as noted by Boring et al. (2015) in their research using a full-scope simulation environment to support control room modernization efforts designed to extend the operating life span of existing NPPs in the U.S.

Discount usability is particularly applicable to the nuclear process control domain because it makes the most effective use of scant participants and opportunities to assess user performance and illuminate shortcomings of the interface. Another important aspect of discount usability is the idea of rapid iterative design, which is necessary to take advantage of the brief time in which operators are actively participating in the usability experiments required to support control room modernization efforts. In these full-scope simulation studies, operators are recruited to participate in week-long experiments (Boring et al., 2012a; 2012b; 2013). Therefore, rapid prototyping in close to real-time is critical to effectively using the scant time operators are onsite and participating in the experiment. Furthermore, measures suffering from technical challenges are swiftly abandoned, since any time spent troubleshooting technical issues detracts from the overall purpose of the usability experiment, which involves gathering useful data to support practical improvements on the interfaces under investigation, as well as providing theoretical results that can inform future design and the research community at large. As a result, the simple and easy-to-administer measures are favored for these high-cost experiments to ensure useful data is collected to generate meaningful and informative conclusions in lieu of complicated but potentially powerful and sensitive measures.

The concept of discount usability extends beyond using the simplest and easiest to administer measures in a typical research setting, such as a full-scope simulator study. Discount usability can be expanded upon by looking towards other simpler research settings that can be used to answer the same or similar questions. For example, academia could potentially perform some of the research that would otherwise be performed at large research firms with full-scope simulation facilities. Academia could use reduced fidelity simulations to examine some of the basic psychological issues encountered by operators in student populations. Certainly, there would be issues with generalizing from reduced fidelity and student populations to power plant control rooms and expert operators, but some of the basic perceptual and cognitive factors, such as SA and attention, may be analogous. In the least, it would be possible to perform initial examinations of psychological constructs in a novice student population, and then verify pertinent results in the expert operators using a full-scope simulation. Microworlds are one promising approach to extend research conducted in full-scope experiments to academic settings with reduced fidelity. Microworlds, with their reduced scope and complexity, afford researchers the ability to more easily tackle process control, human-error, and performance issues without the need for extensive SMEs (Ulrich, Werner, & Boring, 2015; Ulrich et al., 2016). Perhaps one of the greatest benefits of extending research into academia is the ability to perform experiments with sufficiently large samples that are capable of yielding statistically significant results.

Full-scope simulation studies are restricted to very small sample sizes consisting of a handful of operators, and therefore, they cannot generate statistically significant results. Following the discount usability approach with microworlds affords using inexpensive and accessible student populations, which supports sufficient sample sizes to yield statistically significant differences (Waern & Cañas, 2003). A subsequent section on the development of the Rancor Microworld, which is the simulation platform used in this current research project, describes the role of microworlds in nuclear process control research in more detail, as well as outline the necessary characteristics for a microworld suited to conduct research with student populations and produce generalizable results that are meaningful for experts and interfaces in the process control domain.

Chapter 4. Attention-Acknowledgement Measure Development

Relevant Attention Constructs and Attention-Acknowledgement Markers

Attention is an important psychological component of the operators' tasks as mentioned in the chapter devoted exclusively to attention. Several important topics of the psychological construct of attention must be considered to ensure the acknowledgement measure is a valid and effective measure of attention. The subsequent sections describe important characteristics of attention and the supporting rationale for key decisions made for designing the attention-acknowledgement marker itself, such as selecting stimuli characteristics and designating target and non-target states.

Selective Attention

The visual acknowledgement measure is based on the concept of selective attention. In real world situations, including the operators' process control task, the visual system is assailed with a constant stream of visual stimuli competing for attention. Selective attention refers to the ability to focus on a stimulus within the environment while ignoring other competing stimuli (Broadbent, 1958; James, 1890). Selective attention allows us to shift our attention and hold it on a particular stimulus of interest. The original work on selective attention was conducted within the auditory domain (Broadbent, 1958, Moray, 1959). This work lead to the filtering model of selective attention, which proposed that unattended stimuli were filtered out and prevented from higher level cognitive processing. Treisman revised the filtering model to account for some physical and attenuated characteristics of the stimuli that do pass through this filter and undergo some additional processing (1960, 1969). Within the visual domain, the saliency model of selective attention was proposed. It posits that selection of visual stimuli is driven by a bottom-up processing of physical characteristics of the stimuli, such as luminance, color, location, orientation, and motion (Itti, Koch, & Niebur, 1998). A saliency map topographically encodes stimuli in terms of their similarity to surrounding stimuli within the visual scene. A stimulus with a high-conspicuity in relation to surrounding stimuli will result in a higher activation of the topographical feature map, which in turn draws attention towards the location of the stimulus. The less-activated stimuli are then inhibited, which further enhances the attentional capture of the conspicuous stimulus. This initial attentional capture effect based on the saliency of a stimulus in relation to the other stimuli present in the visual scene gives rise to the concept of multiple stages of processing associated with visual attention.

Object and Space Mechanisms of Attention

Another important topic to consider for acknowledgement marker development is object-based versus space-based attention. Object- and space-based attention mechanisms provide alternative explanations for how attention is allocated to different regions in space. Space-based theories frame attention as a spotlight that can be directed through the visual scene (Eriksen & Eriksen, 1974; Posner, 1980; Treisman & Gelade, 1980). All elements that fall within the region of the spotlight are concurrently processed. The object-based theory maintains that the elements and their features form contours that segment the visual scene into objects. The concurrent processing occurs for elements that fall within the same contiguous contoured region that comprises that object. The debate continues, with both theories providing supporting evidence, and it appears that the object-based theory may be a special case of the space-based theory such that the two theories are not mutually exclusive, but rather account for different mechanisms of attention (Kravitz & Behrmann, 2011). The implications for object- and space-based attention are directly relevant for the design of the acknowledgement markers because together they raise the question as to whether the acknowledgement markers should be integrated within the interface elements they are associated with (e.g., object-based mechanism engaged) or if it is sufficient to collocate the acknowledgement markers near the interface elements so that they fall within the spotlight of attention (e.g., space-based mechanism engaged). Ultimately, to strive for the most effective measure of attention, a hybrid approach was adopted. This hybrid approach positioned the attention-acknowledgement markers within indicators and components when possible and otherwise positioned the attention-acknowledgement markers as near as possible when the layout of the indicator and component could not accommodate integration within the object.

Attention-Acknowledgement Marker Characteristics

The conspicuity of the acknowledgement markers is particularly important to consider because the acknowledgement markers are intended to measure the distribution of attention and *not* drive attentional focus. Therefore, it is important to design the acknowledgement markers to be subtle such that they go undetected when attention is directed elsewhere. We can take advantage of the acuity distribution across the eye to reduce the conspicuity of the acknowledgement markers (Anderson, Mullen, & Hess, 1991).

Rotation was selected as the feature to designate acknowledgement marker objects in a target-state requiring acknowledgement via a response as opposed to stationary non-rotating-acknowledgement marker objects. Specifically, rotating Gabor patches (Ishai & Sagi, 1995) were selected as the candidate stimulus for the acknowledgement markers because the contrasting frequency bands can be given a high-spatial frequency below the threshold of the retina's peripheral spatial resolution capacity to resolve the contrasting bands within the patch (Anderson, Mullen, & Hess, 1991). It is quite difficult to determine peripherally displayed Gabor patch's orientation, and therefore, the observer cannot determine the Gabor patch is in the target, rotating state. Conversely, when gazing directly at the acknowledgement markers, the rotation or lack of rotation is readily detectable because the foveal region, which is mostly yoked to the locus of attention and gaze location, can resolve the contrasting bands and discriminate between orientations to determine if the patch is rotating. Thus, the Gabor patches demonstrated promise to differentiate between attended and unattended regions of the display. Another more basic stimulus comprised of a solid rectangle was selected to compare to performance on the Gabor patches. The rectangle with its much higher saliency and conspicuity was included to serve as a more easily detectable stimulus, though it is also more detectable within the peripheral region as well, and therefore, could undesirably capture attention.

Consideration was also made for characteristics used to differentiate between the target and non-target state. Numerous characteristics were considered to serve as designators for the target and non-target acknowledgement marker states. For example, color could be used to differentiate between the target and non-target states much the same as traffic lights indicate which flow of traffic is allowed at an intersection (i.e., the intersections state). However, within the domain of nuclear process control, color is not a viable feature dimension for the acknowledgement markers because it is already reserved for many other uses throughout the control room (Ulrich et al., 2012). The excessive use of color in control rooms precludes the addition of yet another color purely for attention-allocation-evaluation purposes. Another feature considered to differentiate the target and non-target states was the shape of the acknowledgement marker itself. For example, the acknowledgement markers could resemble the controls themselves, such as bar indicators that shrink and extend much like the bar indicators already present in the control room. This approach was quickly dismissed since it has the potential to lead to significant confusion in which the operator could mistake the acknowledgement markers for the actual interface indicators and controls. Confusing the attention markers with actual indicators could lead the operator to perform inappropriate actions. Maintaining a clear separation between the acknowledgement marker features and those of the actual interface are important to keep the acknowledgement markers distinct and avoid any potential confusion.

The shift from the target and non-target states also required careful consideration due to the temporal demands for an online measure of attention-allocation. For example, the speed at which some feature changes to designate target versus non-target state could be altered to adjust the conspicuity. Slow changes were initially considered to reduce the saliency of the acknowledgement markers while they transition between states. For example, a gradual shift between white and black hues was considered; however, to maintain subtlety and avoid attentional capture, the time required to make the transition inconspicuously was deemed too long for measuring attention-allocation within sufficiently short discrete time-windows. Therefore, rotational motion was selected as the dimension to differentiate between target and non-target states for the acknowledgement markers. Specifically, rotational motion was selected because it can be adjusted to manipulate the object's saliency and this type of rotational motion is less ubiquitous than linear motion in the control room. Control rooms do use radial gauges that do exhibit rotational motion; however, these are less common than bar gauges. Bar gauges are more suitable for providing emergent properties (i.e., aggregating multiple indicators across a system), and therefore, are the more numerous type of indicator to be found in the control room. Furthermore, rotating Gabor patches are visually dissimilar to the rotating bar gauges, which renders the acknowledgement markers distinct to the other control room indicators and avoids confusion.

Upon detecting any acknowledgement marker in the target state (i.e., rotational movement), the participant was instructed to respond to the target via selection with a mouse. Clicking on an acknowledgement marker with the mouse was a suitable response since participants already use the mouse to control the simulation interface and it was relatively quick since, in theory, participants were using the mouse to click the acknowledgement marker residing in their current locus of attention. Furthermore, the nuclear domain and nuclear operators are extensively experienced with using a mouse to control the interface while performing their job. Additionally, based on interviews, operators reported that they use the mouse as a pointer and place-keeper within the interface while they perform peer checks with other operators. Based on these observations, the visual acknowledgement using a mouse selection has promise as a secondary task that will minimally interfere with the primary process control task as currently performed by operators.

Attention-Acknowledgement Marker Development

Four studies were performed to develop the visual stimuli used as acknowledgement markers for the attention-acknowledgement measure. Each study was intended to address different aspects of attention, as well as different ways in which the participants could provide a measurable response (see Table 4.1). The attention constructs investigated within the studies included the number of acknowledgement markers present in the display and the resulting density of the display. A full-grid of 32 acknowledgement markers created the greatest density, while a partial grid of 12 was examined to determine the effects of attentional resolution and memory (Ulrich, Werner, & Boring, 2015). Two major response patterns were explored, which were immediate selection upon detection via a mouse click or a key-press decision-based response following the trial. In the decision-based response, the participant was required to remember which acknowledgement markers were rotating and provide a yes or no response when probed about a particular acknowledgement marker post-trial. Accuracy was recorded in each study and was defined as correctly identify rotating, or target-state, acknowledgement markers. Reaction-time was measured only in the studies using the immediate selection response format, while reactions times were not measured for the initial two studies using the key-press decision-based response format. A summary of the experimental configurations can be seen in Table 4.1.

Table 4.1 The configuration for each pilot study to examine attention characteristics and response patterns for the detection and acknowledgement of acknowledgement markers.

Study	Matrix	Targets	Stimuli	Response	Accuracy	RT
1	Full	16	Gabors, Rectangle	Decision	✓	✗
2	Partial	6	Gabors, Rectangle	Decision	✓	✗
3	Partial/Full	1	Gabors, Rectangle	Select	✓	✓
4	Full	2	Rectangle	Select	✓	✓

Despite these variations, each of these studies were quite similar in nature in that they all were comprised of a grid of a primary crosshair tracking task and the secondary acknowledgement marker detection task as can be seen in Figure 4.1.

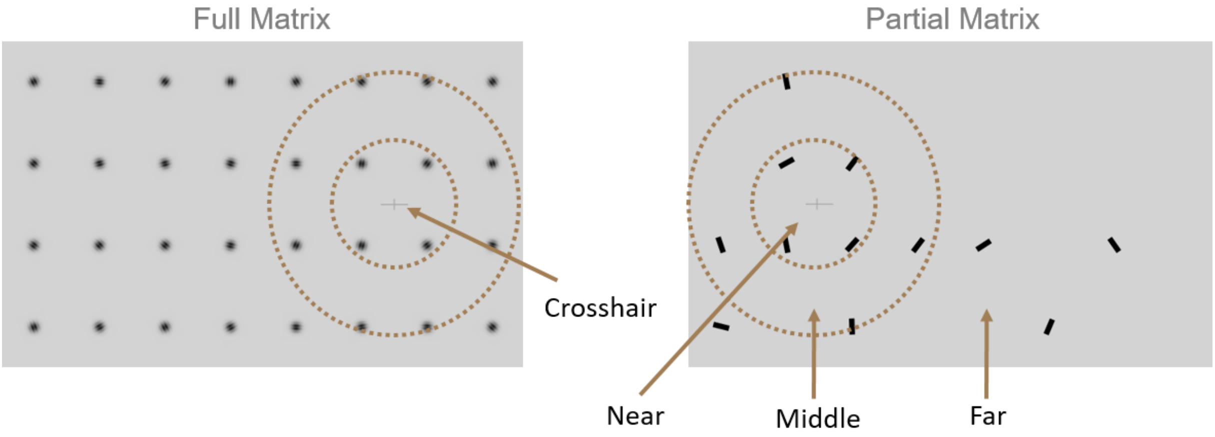


Figure 4.1 Full grid (left) of attention-acknowledgement markers and partial grid (right) of attention-acknowledgement markers used in the measure development. The circles denote increasingly eccentric regions (i.e., near, middle, far) surrounding the locus of attention at the primary crosshair.

The pilot studies used varying numbers of target-state attention markers. Pilot studies one and two use a configuration in which half the acknowledgement markers were in the target-rotating state, while half remained stationary. At the end of the trial, a single acknowledgement marker was highlighted and participants were tasked with responding whether that acknowledgement marker was in the target or non-target state, which yielded an accuracy score. This response format did not lend itself well to an online measure since it required discrete trials, and therefore, the immediate selection format was adopted for the remaining pilot studies. The immediate selection format entailed selecting the first acknowledgement marker detected and ending the trial.

The crosshair tracking task required key presses to maintain the pseudo-randomly moving vertical post at the center-point of the horizontal post of the crosshair. Participants were told their primary task was to keep the crosshair centered, while their secondary task was to identify any rotating-acknowledgement markers. Two different grids of acknowledgement markers were used in the studies. The full-grid configuration display consisted of 32 acknowledgement markers arranged in a 4x8 grid (e.g., the full-grid configuration). The partial-grid configuration display consisted of 16 acknowledgement markers that were randomly positioned within the 32 possible positions in the 4x8 grid. The crosshair task was always positioned horizontally between the second and third rows of the acknowledgement markers grid. This central positioning was used to fixate participants' attention towards the central portion of the displays and support testing the effects of distance on detection accuracy and reaction times.

Table 4.2 Hypotheses examined in the acknowledgement marker pilot studies.

Hypothesis	Supported?
Acknowledgement markers near the primary task are detected with greater accuracy than distant acknowledgement markers.	Yes
Acknowledgement markers near the primary task elicit shorter detection response times than distant acknowledgement markers.	Yes
Acknowledgement markers detection accuracy at distant acknowledgement markers approximate chance levels of detection accuracy.	Yes
Acknowledgement markers detection accuracy are sufficiently sensitive to differentiate attended and unattended regions of the display.	Yes
Acknowledgement marker detection accuracy and response times correlate with eye tracking metrics: a. Undetected acknowledgement markers register fewer fixations as measured via eye tracking. b. Detected acknowledgement markers exhibit more fixations prior to their response. c. Manipulation Check: The crosshair task exhibits more fixations than acknowledgement markers to ensure participants engaged in the primary task appropriately.	Yes

The pilot studies and their hypotheses (see Table 4.2) were intended to answer two primary questions. First, can participants complete both a primary and secondary task without being overwhelmed as evidenced by poor accuracy detection and long reaction times? Second, what features support good detectability when the markers fall within the attended region, but otherwise do not capture attention when it is focused elsewhere? The initial pilot study selected two Gabor patches of different spatial frequencies and a solid rectangle to serve as the acknowledgement marker stimuli. The Gabor patches were each 100x100 pixels, with a light grey background color and a black foreground color with spatial frequencies of 0.3 or 0.6 cycles per pixel. A 100x25 pixel bar was included to serve as a salient alternative acknowledgement marker object to the Gabor patches. Interestingly, the Gabor patches proved too difficult to detect as evidenced by significantly longer reaction times and poorer accuracy across all distances than the rectangle acknowledgement markers. In pilot study three, a single target marker was presented at varying distances within a partial matrix of stationary non-target markers as can be seen in Figure 4.1.

The Gabor patch acknowledgement markers were at best accurately acknowledged at 90% accuracy at near distances with that accuracy dropping to approximately 60% at farther distances. However, participants accurately acknowledged the rectangle acknowledgement markers almost perfectly at near distances and maintained high-accuracy around 90% even at far distances (see Figure 4.2 for detection accuracy data from pilot study three). The rectangle acknowledgement markers demonstrated good performance with faster reaction times and high-levels of accuracy when located near the primary task and significantly slower reaction times and chance levels of accuracy for detection when in far regions (see Figure 4.3 for reaction time data from pilot study three). Thus, the rectangle stimulus was selected to serve as the acknowledgement marker stimulus for the final pilot study and was ultimately adopted for the attention evaluations within the microworld studies.

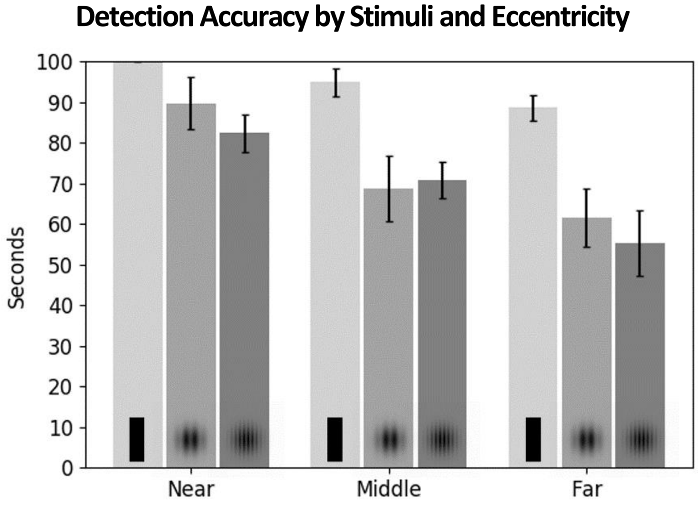


Figure 4.2 Detection accuracy for each stimulus type and at different eccentricities. The accuracy levels of the two Gabor patches tested were too low to be used reliably as a stimulus for the acknowledgement markers, while the rectangle afforded good detectability. Error bars indicate standard error for each mean.

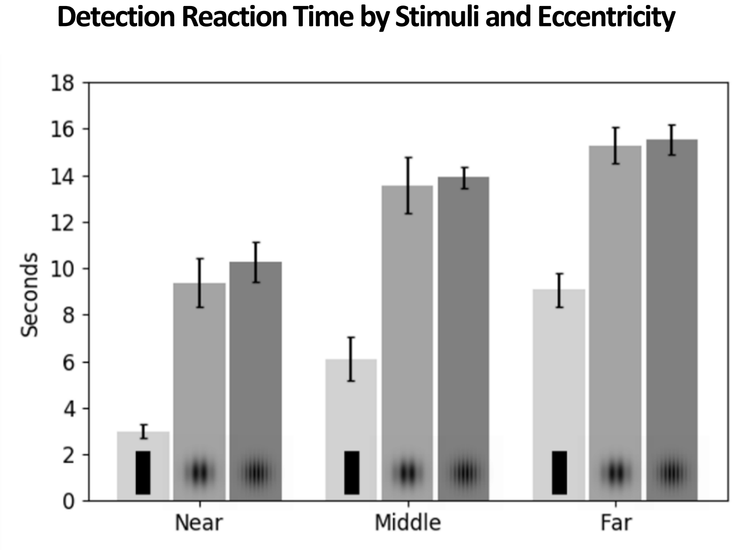


Figure 4.3 Response times in seconds for each stimulus type and at different eccentricities. The rectangle stimulus demonstrated the shortest reactions times while both Gabor patches demonstrated significantly higher reaction times at each eccentricity. As eccentricity increased, the reaction time increased. Error bars indicate standard error for each mean.

Two different response formats were investigated. The first format entailed observing the grid of acknowledgement markers to identify target and non-target states. Participants then had to remember which were stationary and which were in motion. At the end of the trial, a single acknowledgement marker was probed, which required the participant to respond affirmatively if the probed acknowledgement marker had been in the rotating-target state. This response was considered because it generated little response interference with the primary crosshair tracking task since the response itself was made after the trial had commenced. However, the performance was poor as indicated by low-rates of acknowledging probed acknowledgement markers even for acknowledgement markers near the primary task. This response format was also deemed not compatible with a continuous task, such as the process control task. The simulation would have to be broken apart into discrete trials and frozen to allow for a recall-based response, and therefore, this approach was discarded.

Instead, the continuous monitoring and immediate response to the rotating-target state acknowledgement marker upon detection showed more promise. This response approach allowed for the measurement of reaction times, as well as accuracy, which can both be used as indices of attention-allocation. The immediate selection via a mouse response was deemed adequate since participants demonstrated good accuracy performance (i.e., correctly selected the target) and avoided incorrectly selecting non-targets during the pilot studies. Additionally, the immediate selection with a mouse integrates well with the microworld simulation because the simulation also requires the mouse to manipulate interface controls.

The pattern of decline in reaction times and accuracy was sufficient to differentiate between attended and unattended locations as evidenced by the significant increase in reaction times and reduction in accuracy as the acknowledgement markers distance from the primary crosshair task increased. To further validate the use of the rectangle marker stimulus to serve as a measure of attention, a fourth pilot study was performed. In this study, two rectangle marker stimuli were in the active state and positioned at different eccentricities to determine whether participants consistently selected the closer target over the more distant target, as would be expected since the locus of attention fixated on the crosshair task should make closer targets more salient and detectable over more distant targets. Indeed, this pattern of results was observed with participants consistently selecting the nearer of the two targets as can be seen in Figure 4.4.

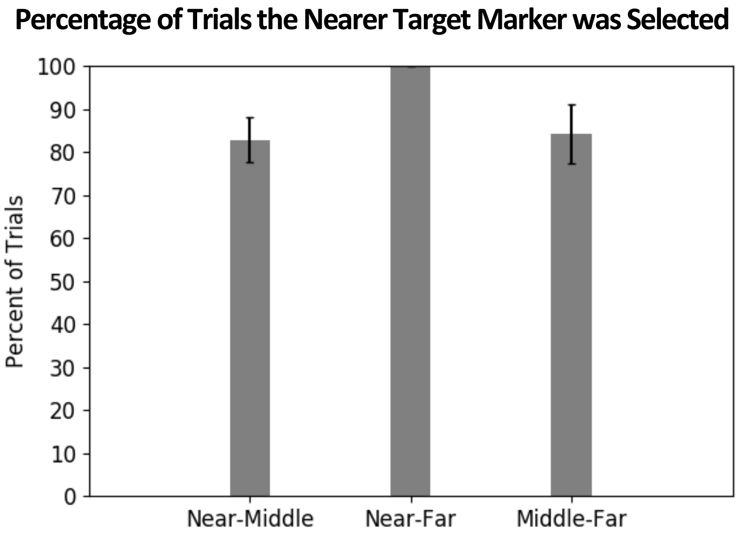


Figure 4.4 Proportion of trials in which participants selected the rectangle marker nearer to the primary crosshair task over a second rectangle marker located more distally. Participants selected the nearer target marker consistently at high percentages over the more distant target marker. Error bars indicate standard error for each mean.

Furthermore, analysis of the eye tracking data in the fourth pilot study was used to validate that acknowledgement markers selected were in fact fixated upon prior to their selection and that acknowledgement markers that were not selected were in fact not fixated upon during the trial. The eye tracking data also served to characterize the interference by the secondary marker-acknowledgement task. As can be seen in Figure 4.5, the unacknowledged target was fixated upon in 10% or less of the trials indicating that participants were not attending to the more distant target marker. As distance from the crosshair task increased, the percentage of trials in which the unattended target was fixated upon declines as expected. Additionally, participants fixated on the crosshair tasks in the majority of trials when the target marker was located near the crosshair, which indicates that the near markers demonstrate minimal attentional interference. Once the acknowledgement task requires participants to begin searching for target marker when the nearest is positioned farther away, the attention shifts away from the crosshair task and more interference is generated. This provides evidence that collocating the marker with the object attention is allocated towards generates a minimal amount of attentional interference and validates the use of the marker to assess attention. Together, the findings from these pilot studies provide evidence that the rectangle acknowledgement marker detection and the immediate selection via a mouse-based response functioned as intended and can serve as a measure of attention-allocation with minimal interference.

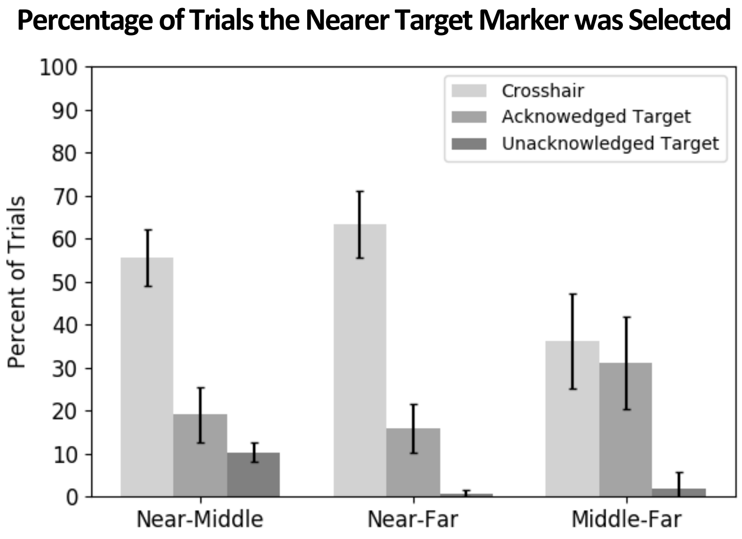

Percentage of Trials the Nearer Target Marker was Selected

Figure 4.5 The percentage of trials in which participants fixated upon the different stimuli during each trial. In general participants fixated upon the crosshair task during most of the trials when one of the target markers was located near the primary crosshair task. The unacknowledged target was fixated upon on few trials, while the acknowledged target was fixated upon on a greater percentage of trials.

Chapter 5. Rancor Microworld Development

The Rancor Microworld is certainly not the first process control microworld simulation environment developed. Other platforms exist, such as DURESS and its descendant DURESS-II, which contain a process flow simulation (Vicente & Pawlak, 1994). DURESS served as an effective tool to evaluate ecological interface designs within the context of human factors (Vicente & Pawlak, 1994; Vicente, Chistoffersen, & Pereklita, 1995; Vicente et al., 1996). Unfortunately, DURESS is somewhat outdated and has not been maintained beyond its second iteration, DURESS-II. A more recent generic process control microworld platform was proposed and developed by Dyre et al. (2013). This tool provided a flexible framework to build simulations of processes tailored to specific experimental questions. The flexibility this tool affords is advantageous, but the Rancor Microworld adopted a more specific design targeting attention and SA psychological constructs. Additional details concerning this philosophy for a more specific design can be found in a subsequent section on the philosophy and technical considerations used for the Rancor Microworld development.

Nuclear reactors in the U.S. adhere to one of two basic designs: boiling water reactors or pressurized water reactors. The primary difference between these two designs is the inclusion of a secondary loop that acts as a further barrier between the radioactive coolant and the outside world. In a boiling water reactor, the coolant itself undergoes a phase change to convert water into steam, which is then used to turn a turbine. In a pressurized water reactor, the primary coolant is kept separate at a higher pressure in the liquid phase within the primary coolant loop. The primary coolant is then used to heat a secondary loop that goes through the phase change to convert water to steam and provide the energy necessary to turn the turbine and create electricity. Of these two designs, the pressurized water reactor design comprises 66 of the 100 operating reactors in the U.S. (U.S. NRC, 2017). The Rancor Microworld simulates a simplified pressurized water nuclear reactor process comprised of a nuclear reactor core that provides the heat source for a gamified water-based Rankine cycle simulation. In fact, the name Rancor, bestowed upon the Microworld, was formed by combining the first part of the term *Rankine* with an abbreviation of the term *core* to denote a nuclear reactor core as the boiler within the Rankine cycle. The water-based Rankine cycle is a mathematical model describing the energy and associated phase changes of a fluid to vapor to convert the thermal energy from steam into mechanical energy used to spin a turbine and generate electricity. The underlying thermohydraulic simulation in the Rancor Microworld diverges from a general water-based Rankine cycle to simplify the system in accordance with the principles of gamification.

The thermohydraulics of the simulation followed a gamified Rankine cycle that resembles the design of a small modular reactor. The model follows basic physics, such as conservation of mass and enthalpy. The core coolant heating and enthalpy losses as the coolant flow through the primary loop rely on difference equations and are parameterized using realistic units. The state change and flow dynamics are largely gamified with little to no fidelity beyond conserving mass. On the primary side, natural circulation of the primary coolant increases monotonically with temperature. Turning on reactor coolant pumps increases flow in proportion to the number of coolant pumps running. The steam generators are implemented such that the water level on the secondary side determines the generator's efficiency, such that optimal efficiency occurs when the steam generators are at the 50% level. As the level deviates from 50%, the efficiency decreases, though for the sake of gamification and usability, the reduction in efficiency always remains quite small. The enthalpy transferred between the primary and secondary loops is governed by the flow rate through the primary loop.

Control of the process resides with the participant; however, a system of interlocks and alarms keeps the plant within a well-defined operating envelope. The interlocks ensure the participant cannot move the plant into an unrealistic configuration and keeps the scenario within a specified band to ensure that the participant can always recover without breaking the simulation. For example, the reactor vessel has a high-temperature set-point to automatically SCRAM the reactor to prevent the reactor vessel temperature from rising to an unfeasibly high-level. In addition to placing limitations on specific component values, the interlocks also ensure the participant cannot operate the plant in nonsensical ways, such as attempting to sync to the grid without the turbine at the required 1800 rpm necessary to produce electricity at the standard 60Hz. Without this protective setup of interlocks, operators could move the plant into unrealistic configurations, such as the core temperature being raised to unfeasible temperatures, if the participant controlling the plant did not provide adequate primary or secondary flow. The interlocks also enforce that certain permissives must be met to put the plant into some control conditions. These permissives are linked to annunciators to convey the current state, such as the annunciator that illuminates to denote the turbine is latched. Additionally, the annunciators arranged in panels at the top of the interface notify operators when certain plant indications are outside normal operating boundaries, if safety systems have engaged. Other microworld environments we have developed have incorporated set-point controllers and decision support systems that would notify operators to take action in advance of indications moving past alarm set-points (Ulrich et al., 2014).

The training and practice required to learn to operate the Rancor Microworld is greatly reduced by using the simplified simulation. The time-course for the simulation was also compressed to allow for the participant to interact with the Rancor Microworld in short durations. This gamification principle of compressing the time-span for a process is advantageous because it affords greater opportunity to collect data on the same process, but over a much shorter time-span. Indeed, pilot testing revealed that undergraduate psychology student participants can learn and operate the simulation at desirable competent levels after as little as forty minutes worth of training.

Rancor Microworld Interface

The Rancor Microworld includes a piping and instrumentation diagram (P&ID) graphical depiction of the components within the system located in the centermost position of the interface. The P&ID graphic can be subdivided into the primary and secondary systems, following the same general delineation NPPs use as conventions within their control rooms. The primary loop consists of the nuclear reactor vessel, recirculating coolant pumps, piping, and temperature instrumentation. Rod controls, located below the reactor vessel, allow participants to control the amount of reactivity, and in turn, the amount of heat added to the primary coolant. The recirculating pumps allow participants to circulate the coolant through the primary loop to transfer heat produced in the core to steam generators, which serve as the interface between the primary and secondary loops. The secondary loop consists of the steam generators, turbine, condenser, feedwater pumps, piping, and valves to control the flow of water and steam through the loop. Participants produce steam in the secondary loop and use several valves to control the flow of steam through the turbine to spin a generator and produce electricity. Central to monitoring and controlling the gamified nuclear process control simulation is understanding this relationship.

Rancor Microworld Components

The rancor microsimulation is comprised of interconnected components. Understanding the relationship between these components is necessary for successful operation of the simulation. The thermohydraulic-based components can be divided into two basic interconnected systems, which are the primary and secondary loops (see Figure 5.1). As stated previously, this two-loop configuration resembles the pressurized water reactors that comprise a large proportion of operating nuclear reactors in the U.S. (U.S. NRC, 2017).

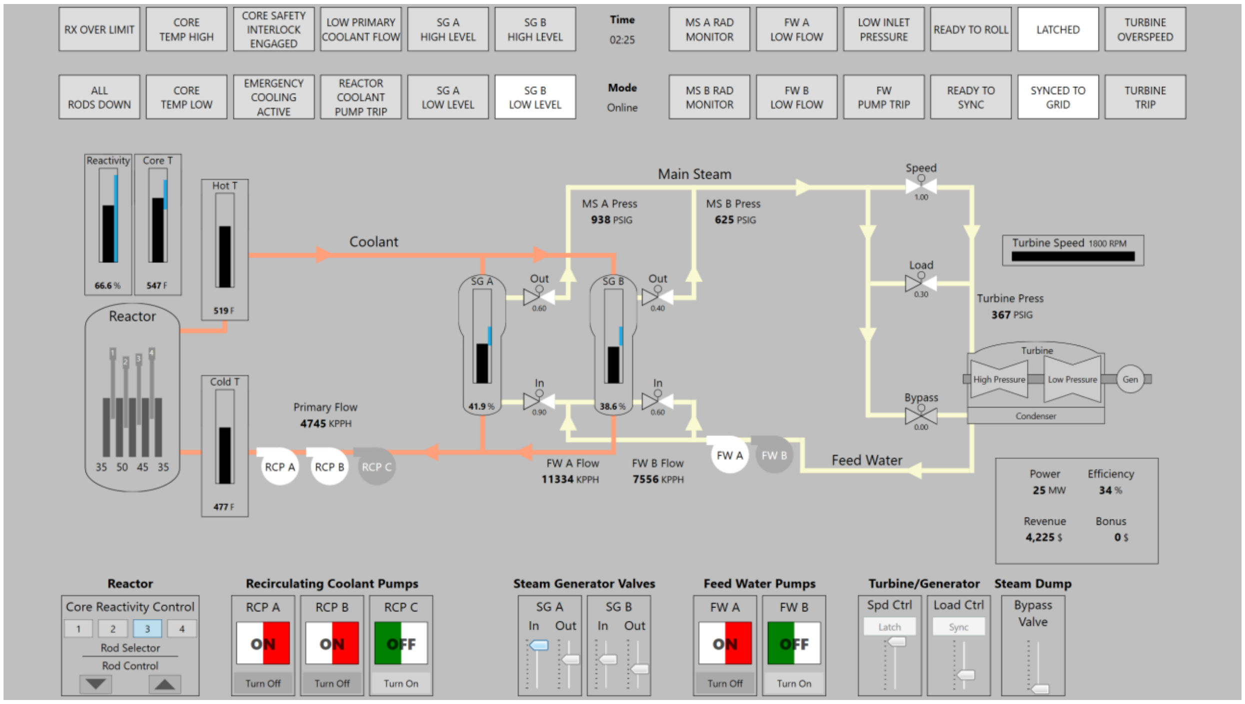


Figure 5.1 Rancor microsimulation depicting the primary loop (orange) containing the reactor vessel and recirculating coolant pumps and secondary loop (cream) containing the steam generators, turbine, condenser, and feedwater pumps.

Primary Loop

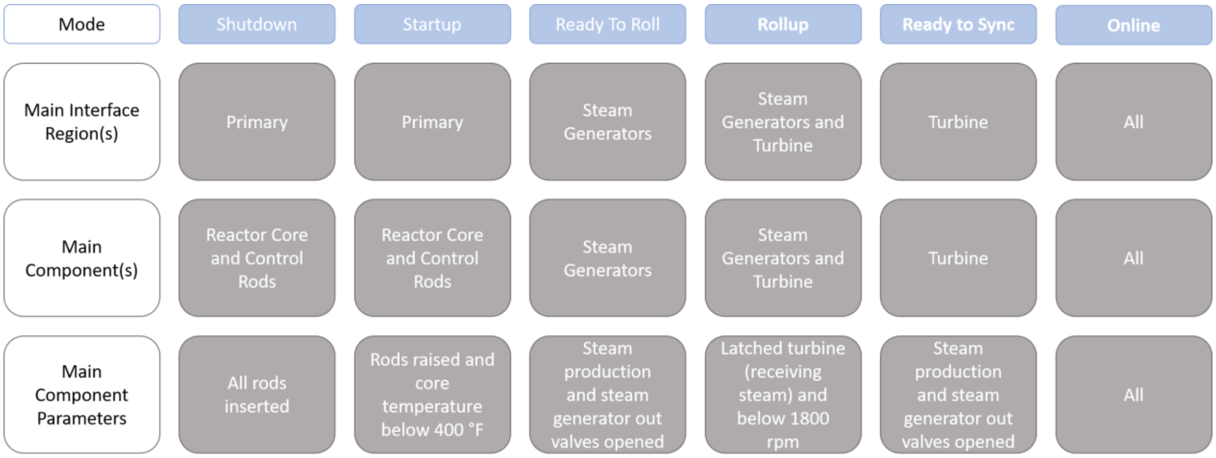
Central to the primary loop is the reactor core vessel, which contains the fuel and control rods. The control rods can be raised to expose the fuel and increase reactivity in the core. As reactivity increases, the temperature of the reactor coolant increases. The reactor vessel serves as the boiler component, or heat supply, within a typical Rankine cycle. The primary loop also contains three recirculating coolant pumps, which provide circulation of the coolant. As temperature in the reactor vessel increases, natural convection increases the coolant flow to a minimal level, but the addition of the recirculating coolant pumps provides a substantial amount of flow. The primary loop interfaces with the secondary loop through the steam generators, which will be described in the next section.

Secondary Loop

The secondary loop contains two primary components: the steam generators and the turbine. The primary reactor coolant flows through the steam generators in an isolated section and transfers its heat to the secondary loop. The outflow of steam can be adjusted by the steam generator out-valves and the inflow of water to replenish the vaporized water in the steam generators can be adjusted by manipulating the in-valves. The steam generated in the steam generators flows through piping towards the turbine and condenser. There are three valves located along the main steam path as it exits the steam generators. The speed control valve provides enough steam to allow the turbine to ramp up to the necessary 1800 rpm required for producing electricity at 60 Hz for the power grid. The load control valve provides an additional and larger flow path for steam to the turbine. The load control valve allows enough steam to provide sufficient torque to the turbine to turn a generator and produce electricity, all while maintaining the turbine at 1800 rpm. A third valve, the bypass valve, is used to divert steam around the turbine in the event that excessive heat has accumulated in the system and must be shed to reduce the amount of thermal energy within both the primary and secondary loops. The secondary loop also contains two feedwater pumps, which produce sufficient flow and positive pressure to replenish the water vaporized into steam within the steam generators and maintain the optimal amount of water-level within the steam generators.

Rancor Microworld Modes

The Rancor simulation has six different modes of operation, which represent different configurations of the plant necessary to achieve sub-goals towards achieving the overall power production goal (see Figure 5.2). The first mode, “shutdown,” represents the plant in the initial state at the beginning of the trial. In this mode, the control rods are fully seated in the reactor and no heat is being produced. The primary loop is initiated at 200 degrees Fahrenheit, but will decay to ambient room temperature if the control rods are not raised to create reactivity and heat the primary loop coolant as it flows through the core. The “startup” mode is defined as the as the reactor core vessel temperature between 200 and 400 degrees Fahrenheit at 10% or greater reactivity. Heating the primary loop sufficiently to support steam production in the steam generators is the primary goal associated with the startup mode. During this mode, participants must monitor the reactivity and core vessel temperature to ensure that the primary loop temperature is increasing at a slow but steady pace. Additionally, participants must power on at least one recirculating coolant pump to provide circulation of the primary coolant through the primary loop. The “ready-to-role” mode is defined by the availability of steam as indicated by the main steam indicators A and B showing steam pressure and the reactor temperature exceeding 400 degrees Fahrenheit. Furthermore, the latch button becomes active to allow participants to latch the turbine. Latching the turbine transitions the plant to the turbine “rollup” mode, which is defined primarily by the speed control valve in an open position. Depressing the latch button “cracks” the valve, which simply means the valve is opened 10%. At this point, the participant can then take control of the speed control valve and move its position to 100%. The valve position regulates the amount of steam flow to the turbine, and in turn, dictates the speed at which the turbine is rotating. Once the turbine has reached 1800 rpm, the plant then enters “ready-to-sync” mode. At this point, the plant is configured to begin power production. Once the load control valve sync button is depressed, the load control valve opens to the 10% position and the plant has entered the “online” mode of operation.



Mode	Shutdown	Startup	Ready To Roll	Rollup	Ready to Sync	Online
Main Interface Region(s)	Primary	Primary	Steam Generators	Steam Generators and Turbine	Turbine	All
Main Component(s)	Reactor Core and Control Rods	Reactor Core and Control Rods	Steam Generators	Steam Generators and Turbine	Turbine	All
Main Component Parameters	All rods inserted	Rods raised and core temperature below 400 °F	Steam production and steam generator out valves opened	Latched turbine (receiving steam) and below 1800 rpm	Steam production and steam generator out valves opened	All

Figure 5.2 The six modes of the Rancor Microworld with the primary components associated with each mode.

Rancor Development Philosophy and Technical Considerations

Microworlds are fundamentally intended to serve as research tools to examine theoretical and practical concepts related to process control. As a research tool, the fundamental objective of the microworld is to extract meaningful and useful empirical data. All other objectives are secondary. Prior design efforts of the Rancor Microworld faltered due to overly ambitious attempts to create a simulation with high-fidelity physics. Ultimately, these high-fidelity versions were discarded since they were overly and unnecessarily complicated. Good physics are inherently a necessity to create a believable simulation, but perfect physicals are unattainable and unnecessary given the scope of this research. Indeed, high-fidelity physics simulations, such as the Reactor Excursion and Leak Analysis Program (RELAP) developed at INL, is comprised of hundreds of thousands of lines of Fortran code created by hundreds of highly experienced experts (Fletcher & Shultz, 1995). This level of fidelity is necessary for RELAP to accurately predict system dynamics for safety purposes; however, when performing psychological studies, a high-fidelity simulation is far beyond the scope of what is necessary for a basic process control system capable of generating meaningful and generalizable results.

Another issue with high-fidelity physical physics concerns the ability to mimic some of the automated control functionality found within the nuclear control room. Many aspects of nuclear process control are operated via control automation in which the operator inputs a set-point and the system automatically maintains the process within a predefined process parameter. The simplified physics of the simulation and the reduced complexity makes it much easier to implement the necessary automatic controls. The simplified physics make it easier to cheat the system and provide the automatic actions to support the same types of interactions found in an actual control room without implementing virtual programmable logic controllers that are found within the physical plant system.

Collectively, these low-fidelity simplifications of the nuclear process control simulation comprising the Rancor Microworld adhere to the concept of gamification. Gamification refers to the use of game design elements in a non-game context (Deterding et al., 2011). As with video games, the virtual space comprising the simulation follows simplified rules that depart from the real world. Similar to a video game, a set of rules and constraints govern the simulation, but the concept of gamification also takes a step farther in the Rancor Microworld. To enhance the engagement of the participants, the principles of gamification were used to incentivize the participants to achieve their optimal performance. A score is displayed for the participant's total revenue, much like the high-score on a video game is displayed, to encourage the participants to push themselves towards better performance.

The microworld was also developed in line with the concept of pragmatic flexibility. The following section contains text from an article published by the author of this dissertation (Ulrich et al.., in press):

> *Pragmatic Flexibility*. Over the last several years, we have embraced a pragmatic flexibility approach for the development of our research tools. Pragmatic flexibility asserts that development cannot begin until there is a specific and targeted research question. This is to ensure that the microworld has at least one use case or research question that it can address to make the development efforts worthwhile in the short-term.
>
> As previously discussed, developing for flexibility and genericity without a clear specification leads to a developmental quagmire in which the development becomes bogged down in striving to make the environment suitable for a diverse range of applications at the cost of being able to address a tangible research question in support of an immediate research need.

Flexibility and generalizability are not inherently bad, they are desirable attributes of software, but they carry an opportunity cost. The additional complexity imposed by new features may escalate in a non-linear manner and quickly become an unnecessary burden.

The Agile development approach is a natural fit for implementing pragmatic flexibility. Agile development is a rapid iterative approach that focuses on implementing features or use-cases in short development cycles (sprints). This works well for microworlds because typically these applications are small in scope and only have to operate in controlled environments for limited periods of time. This greatly reduces the testing and validation compared to having to deploy in a production environment.

Windows Presentation Framework (WPF) is Microsoft's default Windows application framework. The WPF framework has been used to develop interfaces in full-scope simulation efforts involving usability evaluations conducted at the Human Systems Simulation Laboratory at the Idaho National Laboratory (Lew, Boring, & Ulrich, 2014; Boring, Lew, & Ulrich, 2017). WPF leverages Visual Studio and the .NET framework. Visual Studio provides a WYSIWIG for laying out application interfaces. The primary benefit from our adoption is that the modularity of the framework in conjunction with using a Model - View - View - Model (MVVM) architecture is that it allows custom control libraries to be used without modification in microworlds and full-scope HMI prototypes (Lew, Boring, & Ulrich, 2014; Boring, Lew, & Ulrich, 2017). In microworld environments, the object oriented models can be developed in .NET.

Chapter 6. Rancor Microworld Studies

After completing development of the Rancor Microworld and evaluating the attentional characteristics of the acknowledgement markers, the markers were imbedded within the Microworld interface and evaluated as a novel method to assess attention-allocation, and in turn, provide a quantification of Level II SA. The evaluation of the acknowledgement markers within the Rancor Microworld is quite complex due to numerous variables and their corresponding hypotheses. As such, the evaluative studies and their specific hypotheses, analyses, and conclusions are broken into different sections to aid in comprehension. A total of three studies were performed. The first consisted of an evaluation using undergraduate psychology students with no prior experience in nuclear process control or in using the Rancor Microworld. The goal of the first study was to determine if participants could interact with the Microworld and form a baseline characterization of performance, workload, attention, and SA measures. The second study examined a subset of the participants from the original study to further assess performance with increased experience in using the Rancor Microworld and assess impacts of the acknowledgement marker secondary task intrusion and SA probe administration intrusion on primary task performance. The goal of this second study was to examine how participants improved as they gained more experience with the Microworld. The third study was performed to examine performance differences between students and steam plant operators with experience in process control. The goal of the third study was to compare students and operators to determine if the findings of process control Microworld studies can be generalized to expert populations. Furthermore, the third study allowed for an examination of the effectiveness of the acknowledgement marker and SA measures to differentiate between students and operators.

Initial Microworld Evaluation with a Novice Student Population

Since this line of research represents the initial use of the Rancor Microworld, it was important to investigate how novice undergraduate participants performed in controlling the Rancor Microworld simulation. Ultimately, the Rancor Microworld is intended to aid human factors research concerning main control room modernization efforts by serving as a platform for simple, cheap, and easy-to-administer experiments in line with the philosophy of discount usability engineering (Nielsen, 1989). Fundamental to this goal is establishing that novice undergraduates can control the simulation after minimal training, while performing the additional acknowledgement and freeze probe measures of SA. It was predicted that student participants would be able to successfully control the plant as indicated by an overall positive revenue generated and the revenue would increase over subsequent trials as participants gained more exposure and an understanding of the Microworld interface.

Another main objective of this initial evaluation was to compare the attention distributions measured by a traditional measure, such as eye tracking, against the novel acknowledgement marker-acknowledgement measure. In general, the comparisons serves to determine the amount of sensitivity the acknowledgement marker-acknowledgements can achieve in measuring attention-allocation against that of the established eye tracking technique. Specifically, it was predicted that the distribution of fixations within each region of the display (i.e., alarms, primary, steam generators, and turbine) would demonstrate similar patterns to the number of acknowledgements measured by the acknowledgement markers. For example, the proportion of attention measured by fixations in the primary region of the interface was predicted to correlate with the proportion of attention measured by fixations. Additionally, the similarity was expected to be upheld within each mode of operation, since both measures are intended to measure the same construct of attention as the pattern changes between modes of operation due to changing task demands. A detailed examination of the correspondence between the marker-acknowledgements and the eye tracking measurements was also performed. This correspondence extends beyond correlating the two measures, by examining each acknowledgement and determining if an associated fixation in the same interface region occurred. It was predicted that for each acknowledgement marker acknowledged during the trial, an accompanying fixation within the same region would also be captured by the eye tracking recording.

SA is a prerequisite for good performance since participants must have some awareness of the system state in order to perform appropriate manipulations towards achieving a task goal. Thus, it was predicted that participants with better SA would exhibit better performance in the form of total revenue earned by producing electrical power during each trial. Performance was also measured as the time to perform activities within each mode of operation, time to recover from the scripted plant faults, and the error in maintaining reactivity and steam generator levels from their optimal value during the trial.

Better performance was predicted to correlate with errors in the estimation of the parameter values and response accuracy for trend identifications of the parameters across the probe questions.

Another primary objective of this study concerned evaluating a microworld implementation of the freeze probe technique based on the SACRI used in nuclear process control (Hogg et al., 1995). The SACRI was developed for use in full-scope simulations, and therefore, this implementation in a microworld is substantially different. The scope of the simulation in terms of the number of components is greatly reduced, while the time-course for the simulation is substantially more compressed. To the author's knowledge, this is the first evaluation of a freeze probe technique based on the SACRI to be used in a simple microworld simulation over short-duration trials lasting less than ten minutes. The evaluation of the freeze probe technique centered on the degree with which the accuracy of participants' responses to the SA probes relate to their attention allocation. Specifically, the proportion of attention participants allocated to each region of the interface were predicted to correlate with errors in the estimation of the parameter values across the probe questions due to more attention resulting in a greater opportunity to incorporate the parameter value into their SA. Additionally, it was predicted that participants' response accuracy for trend identifications of the parameters across the probe questions within each region would also correlate with the proportion of attention allocated to each region following the same rationale.

Method

Participants. A total of 26 undergraduate psychology students (e.g., 10 female, 16 male) ranging in age from 18 to 47 years (M = 21.27 years, SD = 5.96 years) participated in the study. The undergraduate psychology students were recruited through a research participant pool administration tool, SONA Systems, at the University of Idaho. Students were compensated with course credit.

Protocol. The general experimental protocol included a training session, two guided practice experimental trials, and four experimental trials (see Figure 6.1). An experimenter obtained informed-consent from each participant at the beginning of each data collection session, which lasted about two hours. Participants completed the training during the first 45 minutes of the session, and then spent the remaining time going through the experimental trials. After obtaining informed-consent, participants watched a 12-minute instructional video and were encouraged to pause the video at any time they deemed necessary to ask the experimenter to clarify any concepts they found confusing or unclear. After watching the video, the experimenter then positioned the participant in front of a computer at an appropriate position as indicated by the Tobii Calibration display window. Participants were positioned in front of the monitor and the experimenter verified that the participants were positioned within 70 and 40 cm away from the monitor as required for the Tobii X2-60 eye tracking camera to operate effectively. Prior to each trial, the participants completed a nine-position calibration procedure with the provided Tobii Calibration software. Following calibration, the participants then began each trial. In the first trial, the experimenter walked the participant through the procedures to configure the plant for each mode and respond to the scripted turbine or reactor faults. The second trial consisted of an additional practice trial, but the participants could take the lead while the experimenter monitored their performance and aided as necessary.

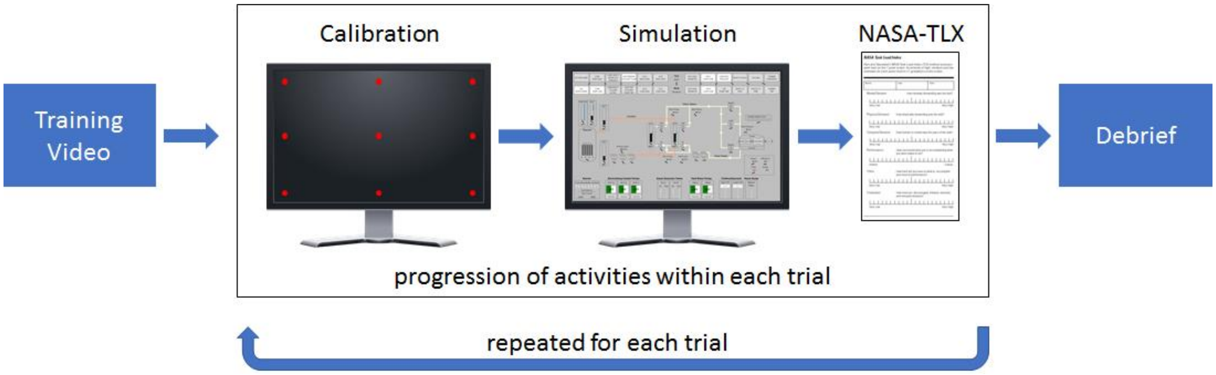


Figure 6.1 General protocol consisted of a training video, trials, and a debrief. Participants first completed two guided practice trials, and then completed four experimental trials for a total of six trials completed during each experimental session.

Training. The training began with a 12-minute instructional video. The video first introduced a general plant model to explain the basic conversion of water into steam through heating in the reactor core vessel, and then how this steam is used to turn a turbine linked to a generator to produce electricity. The video then introduced the Rancor Microworld interface and walked the participant through the alarm panel, the P&ID graphical display, and the controls. Each component and its corresponding controls were explained in terms of how they can be used and what parameters of the process they manipulate. Participants were also informed of the secondary task in which they were instructed to select rotating markers upon detection. The instructions emphasized the monitoring component associated with the secondary tasks to ensure that participants selected markers in a natural fashion while scanning the graphical P&ID display, as opposed to seeking out markers purely to earn a bonus or avoid the warning prompt that occurs if participants do not complete any acknowledgements within every 20 seconds of the trial.

Trial Administration. Each trial was timed to last a total of eight minutes, such that the trial would end at the eight-minute mark, regardless of how the participant performed. Each trial was segmented into five basic time-segments, or epochs, which correspond to the activities the participant performed within each mode of operation (see Figure 6.1). The first epoch consisted of configuring the plant to achieve the latch mode of operation in which the turbine received steam and increased its speed to 1800 rpm. Key tasks that occurred during the first epoch included raising the control rods, powering-on recirculating pumps, adjusting the outflow of steam from the generators, and then latching and fully opening the speed control valve. The first SA probe was administered at a randomly selected time within 20-30 seconds following this epoch. The second epoch was typically shorter in duration and ended when the sync mode of operation had been achieved. During the second epoch, the participant was engaged in monitoring the core temperature and turbine speed, while adjusting the control rods to maintain an appropriate core temperature that was neither rising nor falling rapidly. Some better performing participants began adjusting the level of the steam generators during the second epoch, but the average participant was fully engaged in monitoring the core temperature and turbine speed. Participants selected the sync button upon the turbine reaching its full speed at 1800 rpm. The third epoch consisted of the time between the onset of the fault and the end of the trial. The fault was comprised of either a reactor trip or turbine trip. All control rods fully inserted and reactivity rapidly dropped to zero during the reactor trip. As a direct consequence, the core temperature then began to rapidly drop. Participants were required to identify the issue and raise the control rods to restore reactivity and add heat to the reactor coolant to prevent further complications. If the participant failed to arrest the core temperature reduction and the core temperature dropped below 400 degrees Fahrenheit, then the turbine would also trip and the participant had essentially returned to the initial configuration of the plant at the start of the trial. The turbine trip entailed the turbine becoming unlatched and the participant was then required to latch the turbine and open the speed control valve to the 100% open position. Once the turbine tripped offline, the steam that was previously moving through the turbine no longer had a path to exit the system, which resulted in the core temperature rapidly increasing. The participant could use one of two strategies to arrest the core temperature rising after a turbine trip. The more conservative, but slower, approach was to lower the control rods and reduce core temperature by reducing reactivity within the core. The less conservative approach consisted of using the steam dump valve, which redirected the steam from the main steam lines into the condenser. This process essentially dumped the steam and removed excess heat from the core. Failing to mitigate the core temperature rising could result in excessive core temperature and cause the reactor to trip. If the reactor tripped, the plant had essentially returned to the starting configuration resembling the initial conditions of the simulation at the onset of the trial. The participant was then required to proceed through the general startup procedure to return the plant to the online power-producing state.

Rancor Microworld Attention-Acknowledgement Marker Implementation

The new visual attention-acknowledgement measure was compared to an established eye tracking methodology. To evaluate the utility of the attention-acknowledgement measure as a simple measure to quickly assess attention in complex interfaces, it is necessary to compare its assessment of attention against an established method, such as eye tracking. A total of 38 attention markers were included in the Rancor Microworld interface. The alarms, primary, steam generators, and turbine regions contained 11, 9, 10, and 8 attention markers, respectively (see Figure 6.2).

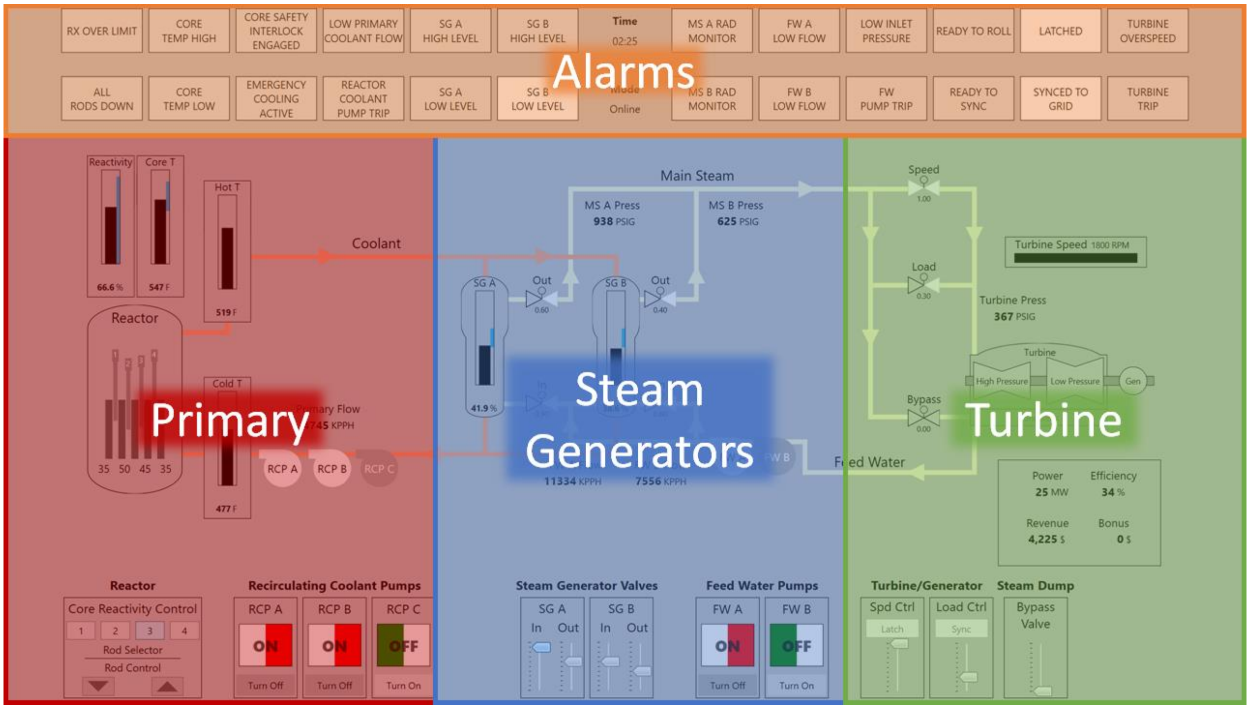


Figure 6.2 The alarm, primary, steam generators, and turbine regions of the screen used to delineate subsections of components with related functions. This color coding convention is adopted throughout this document.

Key components, indicators, and alarms were each assigned an individual acknowledgement marker. Efforts were made to equalize the number of acknowledgement markers within each region to ensure that an equal number were all competing for the operator's attention within each display region. Precisely equalizing the number of markers across all regions was not possible since each region contained different components necessary for the nuclear process; therefore, slight differences in the number of markers in each region was inevitable. Fortunately, the differences in the numbers were small and the distribution of acknowledgement markers across regions was equitable, as can be seen in Figure 6.3.

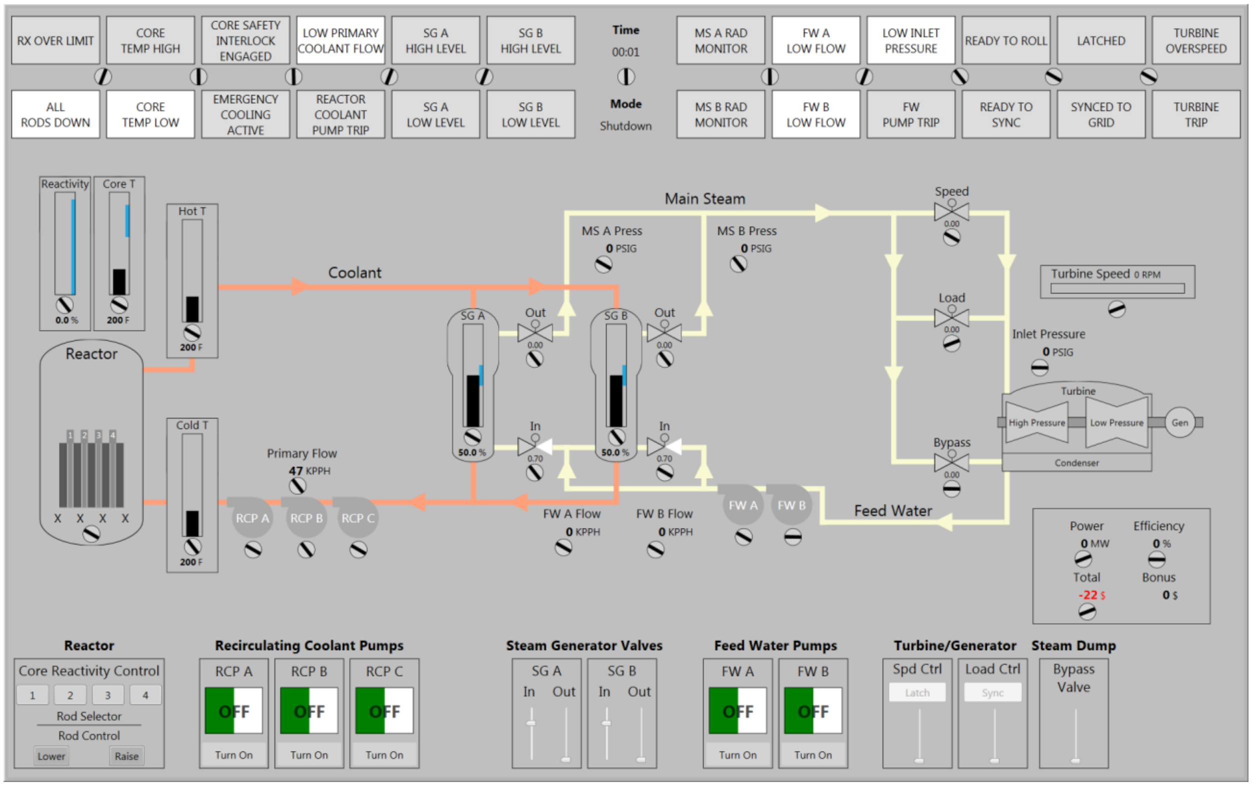


Figure 6.3 Rancor Microworld with attention-acknowledgement markers embedded throughout the display.

The “active” versus “non-active” states of the acknowledgement markers were updated at four second intervals across the trial. During the trial, at any point in time, two-thirds of the marks were in the target-rotating state, while one-third were inactive. The two-thirds proportion for active state acknowledgement markers was selected to provide enough active, rotating-acknowledgement markers so that the participant would not be required to divert their attention away from a region to find a target-state marker. Thus, participants could maintain their natural focus required for the primary task within particular regions of the display without the need to shift focus to fulfil secondary task objectives. Again, it is worth emphasizing that it is important to ensure the acknowledgement markers themselves do not impact where participants direct their attention. Participants can perform the secondary task throughout the interface while the natural movement of their attention transitions across the display as they monitor the components of the nuclear process.

The acknowledgement markers and the eye tracking equipment both measured participants’ distribution of attention as they completed the trial. Since both are measuring the same construct of attention, the two separate attention measures were predicted to demonstrate similar patterns of attention over the entire duration of the trial. Specifically, the proportions of acknowledgements across the alarms, primary, steam generators, and turbine regions of the interface were all predicted to each correlate with the relative proportion of fixations measured via eye tracking within these four regions.

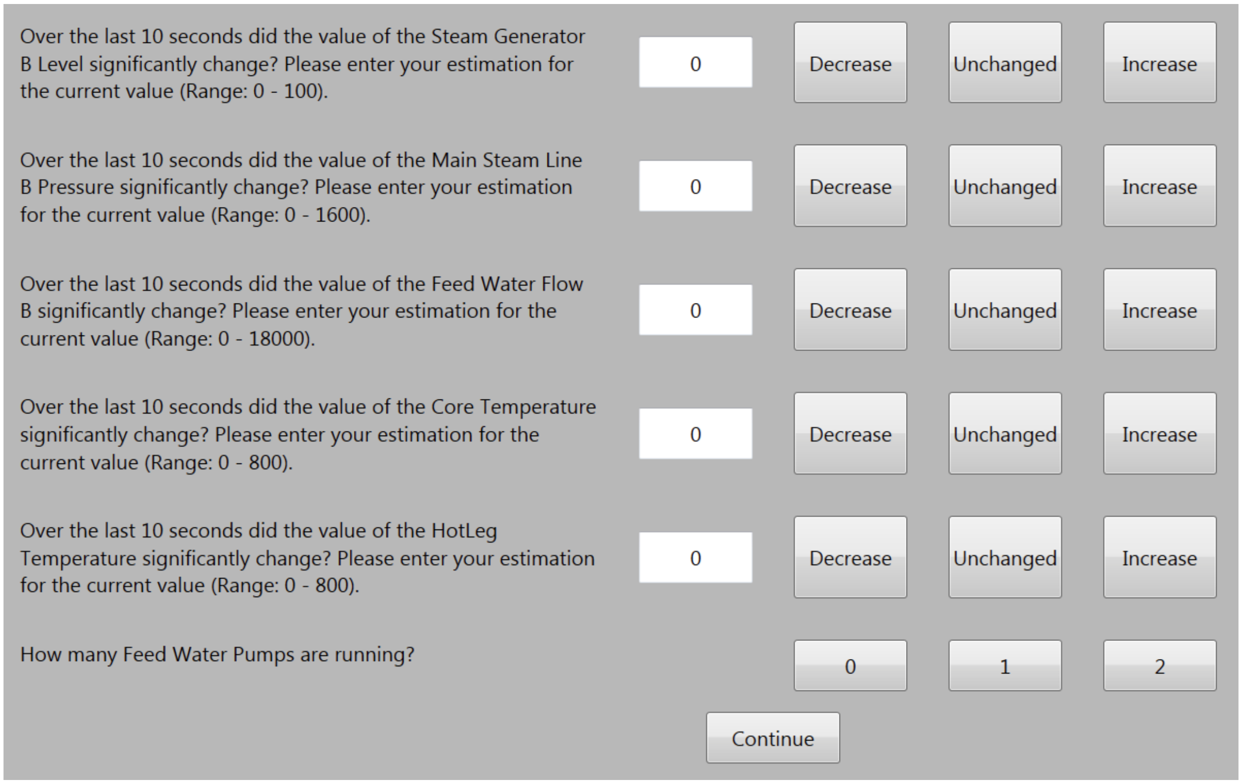
The SEEV model predicts different patterns of attention based on different task demands. For example, the value of an indicator can decrease if it is no longer relevant to the current task, and therefore, less attention should be directed towards it. As such, it was predicted that the modes of operation, with their different tasks, would demonstrate different patterns of attention. These different patterns of attention were predicted to be similar for the fixation and acknowledgement measures within each of these modes of operation, since both are measuring attention. The simulator consists of six modes; however, the first mode, shutdown, represents initial plant conditions. Any actions move the plant into startup mode, and therefore, shutdown mode is of little interest from an analysis perspective. As such, the shutdown mode was considered part of startup for all analyses concerning the modes of operation within the simulation. This resulted in a total of five modes concerning the predictions on the relationship between the markers and eye tracking measures of attention. These mode-specific predictions concern evaluating the utility of the attention-acknowledgement measure to capture nuances of attention distributions at a finer resolution than the general comparison between the two measures across the entire trial.

Rancor Microworld Freeze probe Implementation. Endlsey's freeze probe technique refers to the general procedure of periodically probing the participant concerning key parameters within the situation to assess individuals' SA. The more correctly and accurately the participant can respond to the probes is associated with higher levels of SA. The freeze probe technique requires customized questions targeted to the situation, which in this case is the Rancor Microworld nuclear process control simulation. Endlsey's embodiment of the freeze probe technique, termed SAGAT, formed the basis for the freeze probe measured used in this research. Endsley's SAGAT was designed for use in aviation, but the nature of the probes was similar in that participants were required to estimate and report a specific value at the freeze point. Inspiration was also drawn from the SACRI, which is a version of the freeze probe specifically designed for process control for pressurized light water reactors (Hogg et al., 1995). Since the Rancor Microworld is a simplified simulation of a pressurized light water reactor, the SACRI is an appropriate basis for use as a measure of SA. The SACRI was developed based on an examination of the SAGAT and provide guidance for the development of the SA probes used in the Rancor Microworld. For example, like the SAGAT and SACRI, the probe question content should pertain to concrete physical system parameters, such as temperatures, flow rates, pressure, and levels, as opposed to more abstract thermodynamic concepts, such as enthalpy transfer (Hogg et al., 1995).

The SACRI relies on probes concerning the trend of key parameters as opposed to specific numerical values. Specifically, the SACRI requires participants to report whether the trend of a parameter has remained the same, increased, or decreased over that last few minutes. This is an effective strategy for assessing SA in the control room environment, since plants run in the steady-stable state during most operations, and gradually drifting trends are noteworthy indications more so than specific component values. NPPs are highly complicated systems that are context-dependent. Specific values are not necessarily informative, and therefore, it is necessary to incorporate many parameter values scattered across the control board. As a result, the trend information rises in importance because it provides some aspects of the co-dependence between components. Furthermore, trending values during off-normal and abnormal operating situations reflect dramatic changes in the plant that necessitate operator action, and therefore, trend information is of utmost importance. The microworld is an accelerated simulation, and therefore, the trend information must be assessed during a much smaller timespan. By incorporating aspects of Endsley's SAGAT for specific component values and Hogg et al.'s SACRI for trend information, a comprehensive SA assessment can be performed within the Rancor Microworld.

Performance and SA are related constructs in the sense that good SA is a prerequisite for good performance. Therefore, studies of SA typically attempt to use SA to account for differences in performance observed in the experiment via a correlation between SA and various performance measures. Outside of the original work performed by Hogg et al. (1995) to develop the SACRI method, other work directly using SACRI methodology has not been done, which makes it difficult to determine its validity and relationship to performance. However, the SAGAT has been used more extensively and its ability to correlate with performance has been better quantified. Some studies report that the SAGAT did correlate with performance, while many failed to find a significant correlation. In a series of studies examining whether training in some of the cognitive activities associated with SA, such as attention, can overcome poor SA and subsequent poor performance, the authors found that the SAGAT did not correlate well with performance (O'Brien & O'Hare, 2007). Further examination revealed the SAGAT probes assessing SA at Level I did not significantly correlate with performance, $r(16) = 0.47$, $p > 0.05$, but SAGAT probes assessing SA at Levels II/III did show a significant correlation of $r(16) = 0.63$, $p < 0.01$. The freeze probe questions in this series of experiments assess Levels I and II; therefore, it is predicted that SA will positively correlate in performance, though the inclusion of Level II elements may result in the measure failing to correlate with performance.

During each trial, a total of three freeze probe question sets were completed by participants. A sample administration for a single freeze probe question set can be seen in Figure 6.4. The first two freeze probes were administered following critical events occurring within the simulation. The freeze probes were administered at critical times associated with particular key events within the simulation. The first freeze probe was administered pseudo-randomly between 20 and 30 seconds following the participant successfully configuring the plant for turbine rollup mode. The second probe was administered in the same pseudo-random manner between 20 and 30 seconds following the participant successfully configuring the plant for online mode. The third administration was administered at the end of the trial. The final probe was not linked to a specific event to ensure that a third administration of the freeze probe occurred in the event that the participant could not successfully recover from the scripted fault that occurred with either the turbine or the reactor in each trial. Furthermore, the final administration at the end of the trial ensures that the SA of the participant at the end of the trial is captured, as this provides valuable insight into their final moments while interacting with the simulation. If participants fail to recover from the fault, their lack of SA would also be captured by this final administration. The components included in the pool for random selection was restricted to process values, as opposed to control values. For example, feedwater flow into steam generator A is a potential value to be included, but the value of the valve that controls the flow into the steam generator itself is categorized as a control value, and therefore, is not a potential value that could be included in the probe questions. A total of 16 parameter values comprised the pool of potential probe values, as can be seen in Figure 6.4.



Over the last 10 seconds did the value of the Steam Generator B Level significantly change? Please enter your estimation for the current value (Range: 0 - 100). 0

Over the last 10 seconds did the value of the Main Steam Line B Pressure significantly change? Please enter your estimation for the current value (Range: 0 - 1600). 0

Over the last 10 seconds did the value of the Feed Water Flow B significantly change? Please enter your estimation for the current value (Range: 0 - 18000). 0

Over the last 10 seconds did the value of the Core Temperature significantly change? Please enter your estimation for the current value (Range: 0 - 800). 0

Over the last 10 seconds did the value of the HotLeg Temperature significantly change? Please enter your estimation for the current value (Range: 0 - 800). 0

How many Feed Water Pumps are running?

Figure 6.4 Display of six freeze probes presented to participants while the simulation was frozen. For each probe, participants provided an estimate of a parameter's value and selected a trend for each randomly selected component.

Table 6.1 Parameters included in the freeze probe pool. Six parameters were drawn randomly from this pool for each administration of the freeze probe. Three sets of freeze probes were administered within each trial.

Primary Region	Steam Generator Region	Turbine Region
Reactivity	Steam Generator A Level	Turbine Inlet Valve Pressure
Core Temperature	Steam Generator B Level	Turbine Speed
Hot Leg Temperature	Main Steam Flow A	Megawatts Produced
Cold Leg Temperature	Main Steam Flow B	
Primary Coolant Flow	Feedwater Flow A	
Number of Recirculating Pumps Running	Feedwater Flow B	
	Number of Feedwater Pumps Running	

Each probe question included two parts concerning one of the component parameter values listed in the table above. The first part required participants to enter the value of the component at the freeze point. The question text itself provided the allowable range for a given component (i.e., the reactivity parameter included a range between 0 and 100%). The second part required participants to provide the trend of the component over the last 10 seconds prior to the freeze point. Participants could select either increasing, decreasing, or unchanged as a response option. Probe questions for the recirculating and feedwater pumps were slightly different. Participants were required to provide the number of pumps running only and were not required to indicate a parameter value estimate.

The freeze probes yield an error estimate, which reflects the magnitude difference between estimated parameter value and actual parameter value for the component at the time the probe was administered. Since the components have different scales, the estimates and the true values were first normalized based on the range of possible values for the component prior to calculating the error. This allowed for each probe question to be weighted equally into the overall SA measure. Therefore, each probe resulted in an error value that ranged between zero and one. Performance on each administration of the six probes was then the sum of the error for each probe question.

Performance Measures. Performance on the overall process control task was measured in several different ways. At the most general analysis level, performance for each trial was measured as total revenue was generated. Total revenue is a function of the cost of operating the plant subtracted from the value of electrical power produced, as can be seen in Equation 1.

$$Revenue = \left(MW - 0.5Recirc_{Pumps} - 0.3FW_{Pumps}\right)Rate_{Electric}t\ - 250t - 10000Fine_{Trip} \quad (1)$$

The plant incurs a general $250 operating cost over time, but in addition, the revenue calculation accounts for electricity consumption due to running recirculating and feedwater pumps. Additionally, each reactor trip that occurs during the trial incurs a hefty $10,000 fine. Reactor trips are considered serious violations of safe concepts of operations as evidenced by the following statement concerning the $55,000 fine assessed against Palisades Nuclear Power Plant following a human-error-induced reactor trip, "…the violations reflected significant weaknesses in the planning, communications, and supervision of maintenance work" (U.S. NRC, 1998). To capture the safety culture aspect of the nuclear industry, the $10,000 reactor trip fine provides a strong incentive for participants controlling the simulation to avoid allowing the reactor to trip, as this reflects a substantial portion of the total revenue earned over the course of an eight-minute trial. The critical variable in the revenue equation is the amount of electricity produced, in megawatts, throughout the trial. Participants producing greater electrical output consistently throughout the trial yields an overall higher revenue and indicates good performance.

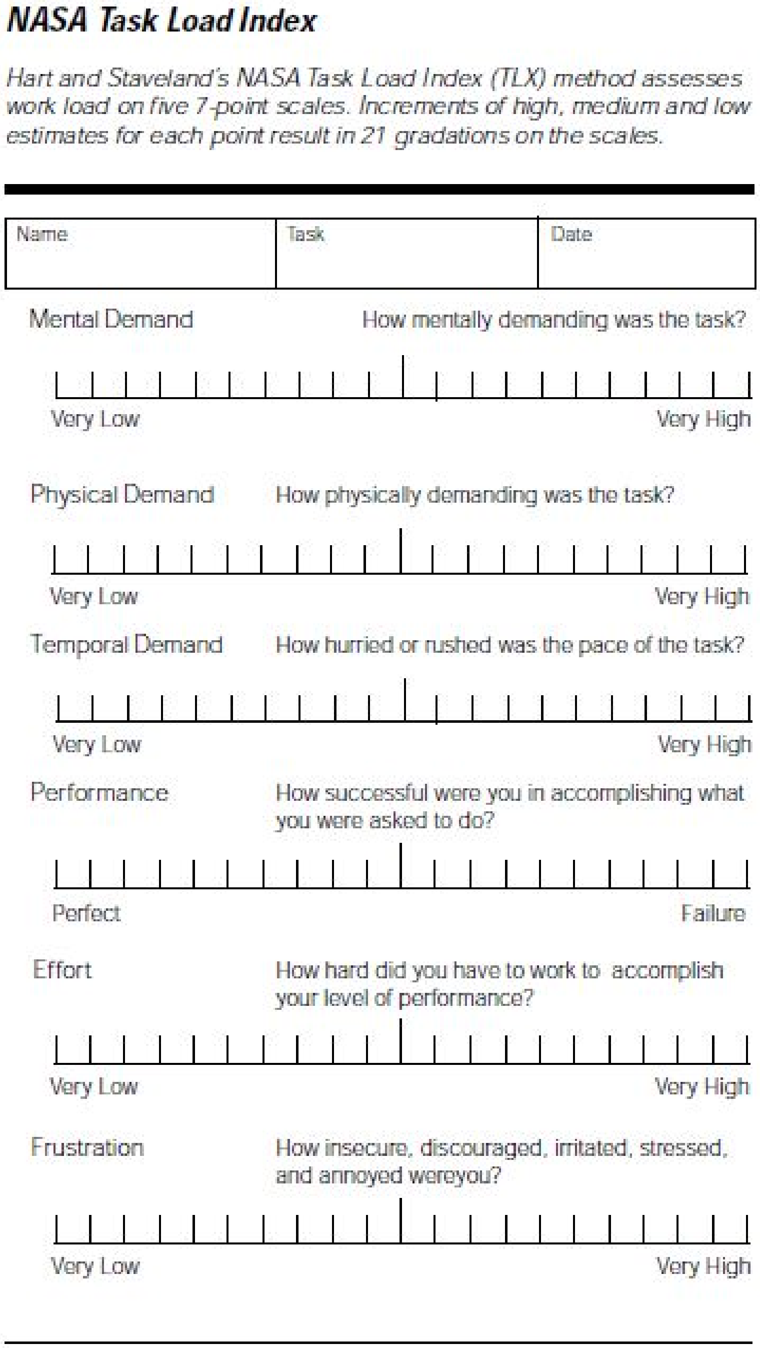
Participants were incentivized to perform the secondary task with a bonus awarded for successfully acknowledging a target-state marker. This bonus was added into the score for the participants; however; the bonus score was not included in the revenue score during the analyses. The bonus does provide a metric of the number of acknowledgements made by participants, but it is simpler to use the count of the numbers as opposed to the actual bonus value. The value of the bonus awarded for successfully acknowledging a marker was set at $75. This value was selected because it was small enough to not drive participants to abandon the primary power production process, but it could also meaningfully improve their score and drive them to complete the secondary marker task.

Performance was also measured as the time taken to configure the plant for each mode of operation. Performance as a measure of time was recorded for the following: latching the turbine, successfully ramping the turbine to 1800 rpm in preparation for syncing, syncing to the grid and producing electrical power, and recovering from the scripted fault. The time to recover from the scripted fault was defined as the time to resync the turbine for a turbine trip fault and the time to return the reactor to its prior reactivity following a reactor trip fault.

Performance was also assessed at the component level for the steam generators. Primarily, the error magnitude from the 50% volume level value of the two steam generators serves as a good performance indicator reflecting how effectively participants managed the steam generators. The steam generators operate most efficiently at the 50% optimal level and participants were instructed to maintain this level throughout the trial. This optimal 50% volume level is independent of plant configuration, and therefore, this level within the steam generators should be maintained regardless of the other tasks being performed. A visual cue highlighting the normal operating range was included in the steam generator graphic to aid participants in maintaining the steam generators at the optimal 50% value.

Subjective Performance Measures. The National Aeronautics and Space Administration (NASA) task load index (NASA-TLX) is a measure of workload used to assess participant performance during or after the completion of a task (Hart & Staveland, 1988). The NASA-TLX has been used extensively in many domains with diverse task types due to its reasonably easy administration and the sensitivity to experimental manipulations impacting workload (Hart, 2006). Indeed, Hart (2006) estimates that over 300 studies have used NASA-TLX, primarily in traffic control and civilian and military aviation. The original NASA-TLX used in this line of research is a paper-and-pencil version of the measure, as can be seen in Figure 6.5; however, there is a software version of the NASA-TLX available for use as well (Cao et al., 2009). The measure is a multidimensional assessment of workload in which participants rank their subjective experience on six different dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration. Participants rank their experience within each dimension across a 21-point open-ended bipolar scale (see Figure 6.5). In the Rancor Microworld evaluation studies, a total of six NASA-TLX's were completed by each participant during each experimental session.

A common variant of the NASA-TLX is the RAW-TLX, which is used in this line of research. RAW-TLX is used because it is simpler and faster to administer since it does not include the participant making pairwise comparisons of the six dimensions to generate relevance weightings for each dimension in relation to the task (Hart, 2006). There is mixed evidence as to whether the diagnostic sensitivity of the RAW-TLX exceeds that of the original NASA-TLX and the author of the measure suggests to pick the version that is most appropriate given the time constraints of the application (Hart, 2006). In the RAW-TLX, the six dimensions are either totaled or averaged to yield an overall score of workload. In this line of research, the average of the six dimensions are reported. To avoid confusion, any reference to the RAW-TLX variant of the measure will be avoided and instead all subsequent references to the measure will simply be in the form of NASA-TLX.



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High

Figure 6.5 Paper-based NASA-TLX completed by participants after each trial of the Rancor Microworld simulation.

In addition to the subjective workload measure, participants completed a debrief questionnaire after each of the six trials. The debrief questionnaire was designed to assess the participant's mental model of the simulated system they interacted with in the trials. A mental model can be defined as, "…deeply ingrained assumptions, generalizations, or even pictures or images that influence how we understand the world and how we take action. Very often, we are not consciously aware of our mental models or the effects they have on our behavior" (Senge, 2006, p. 8). To assess participant's mental models, the debrief form instructed the participants to recreate the graphical section of the interface depicting the two flow loops and the graphical icons representing the components and indicators in a simple drawing. This mental model sketch was then scored similarly to a content analysis. Content analysis entails examining and coding text based on the underlying concepts they represent (Graneheim & Lundman, 2004; Hsieh & Shannon, 2005). Though the sketches are primarily nonverbal, the same content analysis procedure can be used to code the objects of the system the participant represented with their sketch. Raters examined the drawing to determine if each of the components and indicators was included in the participant's sketch of the system to yield an aggregate score reflecting how comprehensively the participant captured the system.

Results and Discussion

Subjective Performance Analysis. At the end of each experimental trial, participants completed a paper-based NASA-TLX form to rate their subject workload. As stated earlier, the NASA-TLX quantifies subjective workload across the following six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration. The raw NASA-TLX scores for student participants can be seen in Figure 6.6. These scores were analyzed using a repeated measures 4 (trial) x 6 (dimension) Analysis of Variance (ANOVA). A significant main effect for trial was observed, *F*(3,69) = 3.36, *p* = 0.02. A post hoc analysis with Fischer's least significant difference (LSD) showed ratings on the fourth trial were significantly lower than trial one, two, and three. Trials one, two, and three were not significantly different from each other.

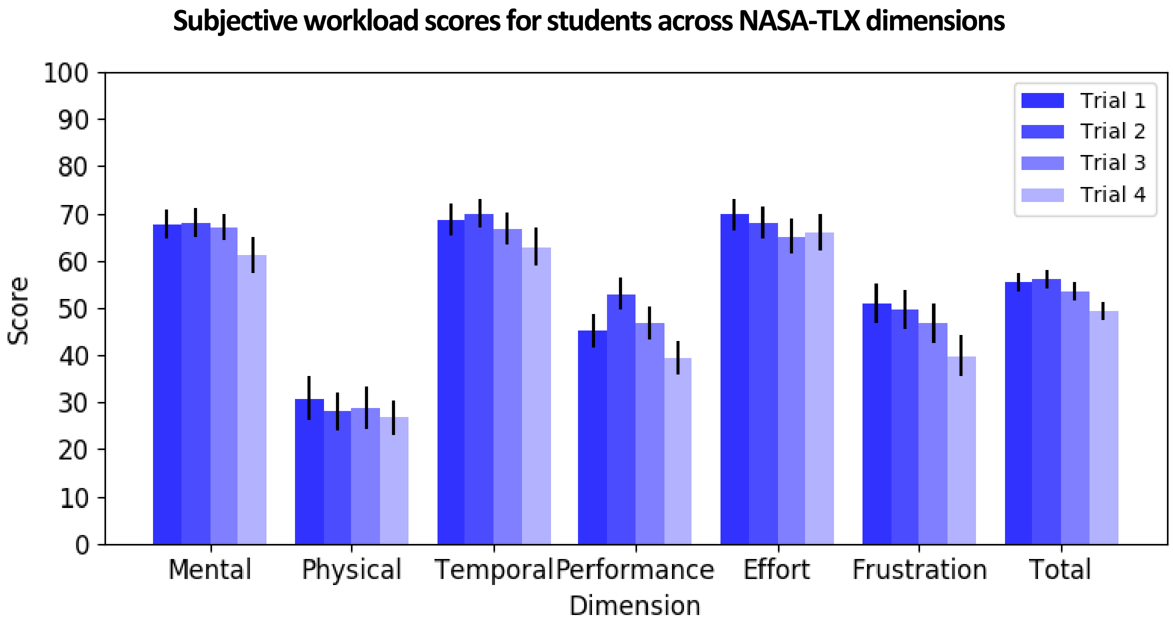


Figure 6.6 Student subjective ratings of workload for each dimension of the NASA-TLX on each of the four experimental trials.

In general, the overall workload of participants assessed by the NASA-TLX yielded scores ranging from 50 to 60%. These scores reflect a moderate amount of workload, which indicates the simulation was deemed engaging and challenged participants to some degree. To put this range of difficulty in perspective, workload reported by nurses performing various activities in an intensive care unit at a hospital yielded scores ranging from 60 to 80% (Hoonakker et al., 2011). These scores reflect the workload in a challenging and stressful environment and are greater than the scores reported by participants using the microworld. Another study examining workload while participants drove a vehicle and interacted with a speech-based email client yielded scores ranging from 30 to 60%, which represents a simple driving task and some verbal distraction (Lee et al., 2001). The scores reported by participants using the microworld are above these values indicating the simulation is more challenging than a typical driving-with-verbal-distraction task.

The overall trend across the trials follows a gradual reduction in ratings for mental demand, temporal demand, and frustration, which is not unexpected since participants gained understanding of the simulation and subsequently were more capable of effecting desired changes. This is further reflected in the debrief form and conversations with the participants in which they commonly reported they were able to successfully recover from the trip events and operate the system effectively. For example, one participant noted, “the first two times I required assistance, after that I was able to figure it out.” Another participant reported they could recover from the scripted trip events, “most of the time, one time I had a lot of trouble lowering the reactor temperature.” Physical demand and effort remained stable and unchanged across the four trials. Interestingly, the performance ratings declined; in part, this may have been attributed to participants experimenting with the simulation as they became more comfortable with it. Upon being questioned post-trial, participants completing trials with a negative revenue commonly replied they attempted to try a new strategy and ended up pushing the system out-of-balance to the point they had difficulty recovering.

Performance Analyses. Central to this initial microworld evaluation was determining whether or not an undergraduate population could learn how to control the simulation and also characterize performance following minimal training in a single experimental session. The criterion for success on a given trial was defined as achieving the online mode of operation and generating an overall positive revenue. There were no instances in which a participant was not able to successfully achieve the online mode of operation. Of the 104 total trials completed during the initial study, participants failed to generate a positive total revenue on only 10 trials. The first and second trials each suffered four of these failures, while the third and fourth trials each only suffered a single failure to generate positive revenue. The overall success-rate based on the initial four trials was 90.38%, which indicates undergraduates were able to successfully learn to control the simulation within a single two-hour training and experimental session.

With increased exposure and practice monitoring and controlling the microworld, participants were expected to demonstrate increased performance on subsequent trials. Participants were explicitly instructed to maximize revenue as the primary goal during the trials, and therefore, revenue was the primary performance indicator. However, two other types of metrics were also used to characterize and examine performance, which included the time to complete activities associated with each mode of operation and the amount of error in controlling specific components in relation to their optimal values over the course of the trial. Furthermore, these three metrics of performance form a cohesive representation, but it should be noted that they are not independent of one another. For example, transitioning through modes of operation faster also tends to result in higher revenue, assuming the participant does not encounter other complications during the trial.

Revenue. The revenue generated during each trial served as the primary and most general metric of performance. The revenue earned by student participants ranged from $4,747 to $71,022 with an average score of $29,967 (*SD* = $17,286). With additional exposure and practice with the microworld, participants were predicted to demonstrate an increase in the total revenue earned on each subsequent trial. Of the six trials participants completed, only the last four were analyzed since the first two were guided-practice trials. Total revenue earned was analyzed across the four experimental trials using a repeated measures ANOVA. A significant main effect for trial on revenue earned was observed, *F*(3,75) = 3.32, *p* = 0.02. A post hoc analysis using Fischer's LSD showed the third and fourth trials were significantly different from the first trial. As can be seen in Figure 6.7, students demonstrated improved revenue on subsequent trials. The increase in revenue reflects improved performance and provides evidence the students were improving their mental model of the simulation throughout the course of the experiment. Furthermore, the total revenue improvement across trials reflects a 44.17% gain from the first experimental trial to the last, which is a substantial improvement.

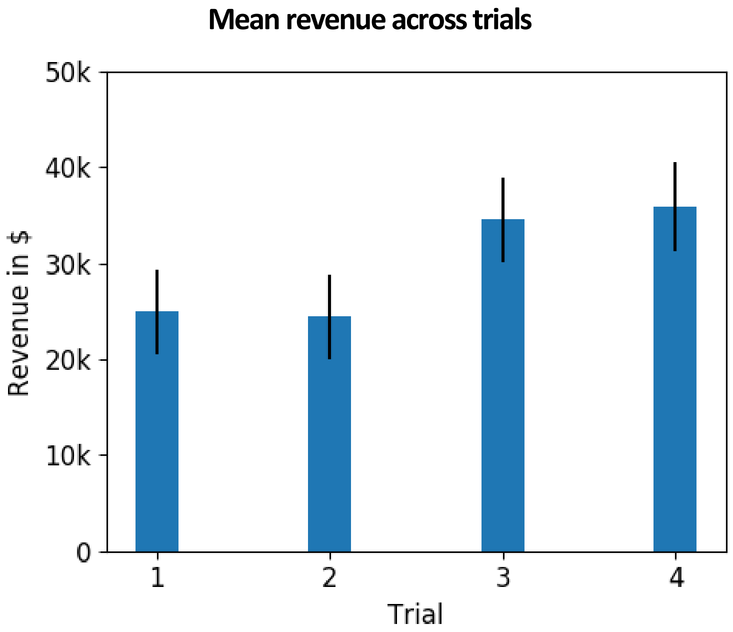


Figure 6.7 Mean total revenue earned by students. Revenue on the third and fourth trial was significantly greater than the first trial. Error bars indicate standard error for each mean.

Time. The average time spent in each mode of operation can be seen in Figure 6.8. Students spent the majority of their time in the online mode of operation, but still spent a considerable amount of time in the startup and rollup modes of operation as well. As can be seen in Figure 6.8, students spent very little time in the ready to roll (*M* = 16.18 s, *SD* = 7.99 s) and ready to sync (*M* = 8.14 s, *SD* = 6.18 s) modes of operation. These two short-duration modes required monitoring, but only consisted of a single action required to transition the system to the next mode; therefore, the timespans were quite short.

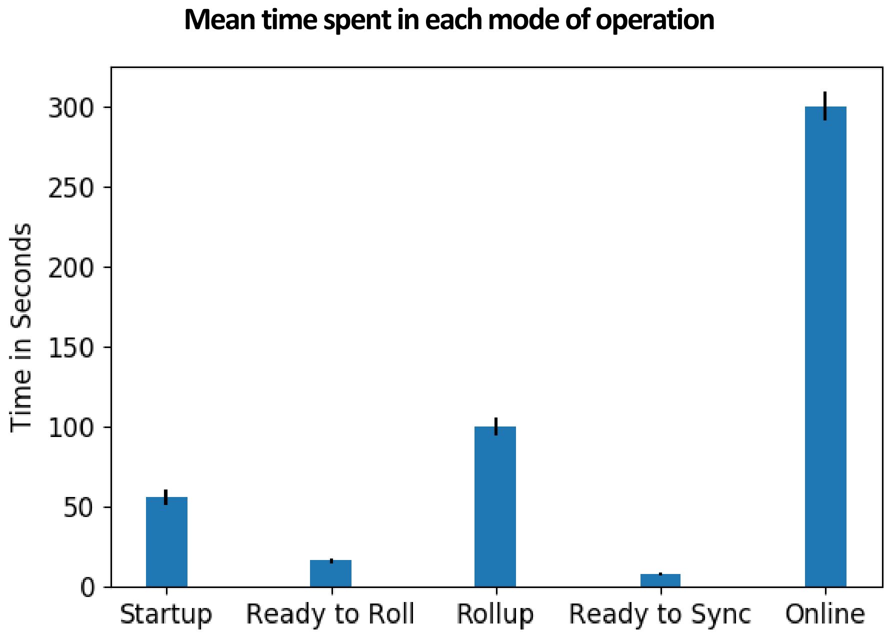


Figure 6.8 Mean time spent within each of the five modes of operation contained within each trial. Each trial lasted a total of 480 seconds (8 minutes). Error bars indicate standard error for each mean.

The time to complete various activities during the trials served as another metric of performance. Manipulating components into more optimal configurations allows for moving through the startup, ready to roll, rollup, and ready to sync modes of operation more quickly than less optimal configurations; therefore, the timing date for each mode captures an aspect of participant performance. As students gained exposure and practice with the microworld simulation, they were predicted to exhibit shorter times to accomplish the activities associated with each of the modes of operation. Specifically, participants were expected to demonstrate shorter times to achieve the turbine rollup mode of operation on each subsequent trial. The time to achieve the turbine rollup mode of operation consists of the elapsed time between the initiation of the trial and configuring the plant to support the latching activity, which defines the boundary between the end of startup and the beginning of the rollup mode of operation. Trial performance as measured by the time to achieve the turbine rollup mode of operation was analyzed across the four trials with a repeated measures ANOVA. No significant main effect for trial on the time to achieve the turbine rollup mode of operation was observed, *F*(3,72) = 1.85, *p* = 0.15. The results suggest that the participants did not meaningfully improve in their ability to quickly move the plant into the turbine rollup mode of operation as seen in Figure 6.9. Another timing performance indicator, syncing the turbine to the grid, represents a significant transitional time-point in which the plant shifts from incurring costs to generating electrical power and profit. To avoid the time to achieve the previously examined rollup mode from confounding performance during this next phase of the process, the elapsed time between turbine rollup and the online mode of operation was used instead of the actual trial time. In line with the previous hypothesis concerning the time to achieve the turbine rollup mode of operation, participants were predicted to spend less time in the interim modes between turbine rollup and the online mode of operation on each subsequent trial. The elapsed time between the rollup and online modes of operation was assessed with a repeated measures ANOVA. No significant main effect for trial on the elapsed time between achieving the rollup and online modes of operation was observed, *F*(3,72) = 0.80, *p* = 0.50. As with the time required to achieve the ready to roll mode of operation, the elapsed time from achieving the ready to roll mode of operation to achieving the online mode of operation did not improve meaningfully over the course of the trials.

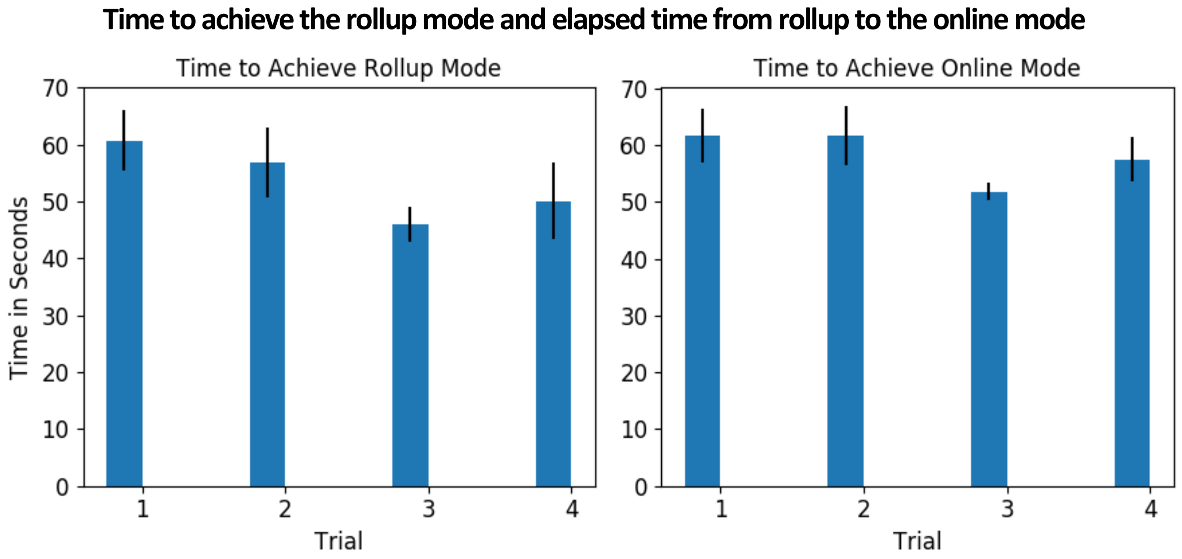


Figure 6.9 The mean elapsed time, in seconds, from the beginning of the trial and the participant achieving the ready to roll mode of operation (left) and the elapsed time from the ready to roll mode of operation and the participant achieving the online mode of operation (right). Error bars indicate standard error for each mean.

In each trial, participants encountered a scripted and unavoidable reactor trip or turbine trip. These two faults were selected because they affect components that are positioned on opposite regions of the interface. The reactor trip requires participants to direct attention to the primary region of the interface, while the turbine trip requires participants to direct attention to the turbine region of the interface. These two faults were selected for several reasons. First, each fault is a salient event marked by visual changes in the display in the form of the control rods all moving to the fully inserted position in the reactor trip or the turbine speed indicator, represented as a horizontal bar meter receding to its zero-point value as turbine speed drops rapidly during the turbine trip. Secondly, the faults were intended to introduce some variability due to differences in the difficulty posed by recovery actions. The time to recover for the turbine fault was predicted to be longer, since the recovery time for the turbine fault is limited by the acceleration rate of the turbine. Furthermore, a failure to respond to the turbine fault in a timely manner can cause a cascading effect in which excess heat builds up in the system and results in a high-temperature-induced reactor fault. The excess heat must be dissipated and the control rods repositioned prior to re-latching the turbine. Because of the cascading effect and the acceleration limitations of the turbine, the time to recover for the turbine fault was predicted to be longer than the recovery time for the reactor fault. This increased recovery time was predicted to translate into decreased total revenue generated during turbine fault condition trials. The order for receiving the turbine and reactor faults was randomized, and therefore, the analysis of the two-fault conditions was collapsed across trials. Two paired-sample t-tests were performed to analyze the time to recover and total revenue earned between the reactor and turbine fault. There was a significant difference in the time to recover between the two fault conditions *t*(24) = 22.62, *p* = 0.00. As predicted, the turbine fault condition exhibited longer recovery times (*M* = 58.06 s, *SD* = 28.37 s) than the reactor fault condition (*M* = 27.79 s, *SD* = 15.18 s). However, contrary to the predicted hypothesis, no significant difference in total revenue was observed between the two conditions *t*(24) = 1.08, *p* = 0.31.

Error in Controlling Components. The average absolute deviations, or error, from optimal values for reactivity in the core and the steam generators' level also served as performance indicators. As with total revenue and timing-based measures of performance, the participants were expected to demonstrate improved performance on subsequent trials.

The amount of revenue generated is governed by the amount of heat added to the system via manipulating reactivity with control rod positions. As such, higher reactivity rates allow for higher total revenue and represent a superior control strategy for operating the microworld simulation. Indeed, this control strategy is analogous to the concept of operations for real world NPPs, since they operate at 100% power to maximize efficiency. Participants were predicted to demonstrate improved performance as indicated by higher levels of reactivity on each subsequent trial. The effect of trial on reactivity was assessed using a repeated measures ANOVA. A significant main effect for trial was observed on average reactivity, *F*(3,75) = 4.42, *p* = 0.01 (see Figure 6.10). A post hoc Fischer's LSD test showed significant differences in reactivity between trial one and trial three and four, which drove the main effect.

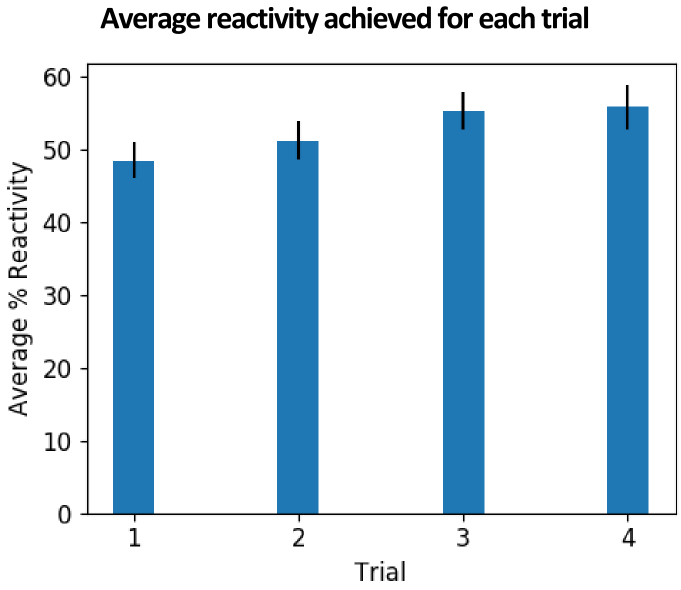


Figure 6.10 Average reactivity for each trial. Higher reactivity represents a better control strategy, which entails adding more heat to the system that is then converted into electrical output and higher total revenue earned. Error bars indicate standard error for each mean.

Better performance monitoring and controlling the steam generators was characterized by smaller error from the optimal 50% value in which participants were instructed to maintain the steam generator level. Participants were explicitly instructed to maintain the steam generator levels at 50%. Additionally, within the microworld simulation mathematical model, the steam generator level imparts a small effect on the efficiency of the plant and electricity produced. Optimal efficiency for producing steam within the steam generators occurs when the steam generators are operating at 50%. The average absolute error from the optimal steam generator level was examined across trials using a repeated measures ANOVA. No significant main effect for trial on steam generator average absolute deviations was observed, $F(3,72) = 2.26$, $p = 0.09$.

In the most general sense, the results suggest that students were capable of controlling a gamified complex process as implemented in the Rancor Microworld simulation. Collectively, the various performance measures also provide evidence for modest improved performance over subsequent trials, indicating that learning occurred during the first experimental session. The students demonstrated increased revenue across trials and improved control strategies as evidenced by higher levels of average reactivity achieved on subsequent trials. The two fault conditions did demonstrate differences in the recovery times they required, with the turbine fault eliciting longer recovery times. However, these recovery times did not translate into significant differences in performance as measured by total revenue. A follow-up study was performed to further examine if performance would improve after additional exposure and practice with the microworld. The results of this follow-up study will be described in detail in a subsequent section.

Situation Awareness. SA was measured with two types of freeze probes. A component value estimate probe type based on Endsley's SAGAT (1995b) and a trend identification probe type based on Hogg's nuclear control domain specific SACRI (1995) were administered to the participants at three time-points throughout the trial. The SAGAT-like component value estimate responses were first normalized within the component to generate an error percentage. This was necessary to compare the magnitude of errors across different components since they operate at different ranges. For example, estimates of reactor core reactivity ranges between 0 and 100%, while estimates of the turbine rpm speed range between 0 and 1800. The SACRI-like trend identification responses were simply the percentage of correct trend identifications for each probe response and its associated component. Freeze probe responses were examined within regions and the time-point, designated as period, at which they were administered. The first freeze probe was administered after achieving the rollup mode of operation, the second after achieving the online mode of operation, and the third at the end of the trial. Each set represents a different system state and activities. It was predicted that the first period would exhibit the best SA scores since at this point only variables within the primary region are fluctuating, which creates a smaller set of components for the participant to monitor. Conversely, the online mode of operation contains all the components fluctuating based on the participants controlling them over the course of the trial and poses the most challenging time-point to interpret, and therefore, the second freeze probe period was predicted to exhibit the worst performance. In regard to region, it was predicted that the primary region would demonstrate the best SA scores, since this region contains the most critical component of the system, the reactor core, and would therefore exhibit the highest proportion of attention throughout each trial. Furthermore, this is the least complicated aspect of the system since it is the first input into the system, and therefore, has fewer reciprocal relationships with other components. As a result of its position within the system with fewer reciprocal relationships, it is less challenging to monitor the primary regions state throughout each trial.

SAGAT-like SA. The average responses for probes in the SAGAT-like component value estimate error and SACRI-like trend identification accuracy measures can be seen in Figure 6.11 and Figure 6.12. A 3 (region) by 3 (period) repeated measures ANOVA was used to assess SA as measured by the component value estimate error probe type within the different time-points and their associated modes of operation. No significant main effect for period $F(2,46) = 0.61$, $p = 0.55$ was found, and therefore, the hypothesis concerning better SA during the first probe administration was not supported. A significant main effect for region $F(2,46) = 17.12$, $p = 0.00$ was found. As predicted, the freeze probe responses within the primary region demonstrated the lowest error rates, denoting better SA (see Figure 6.13). A Fischer's LSD test showed a significant difference between the primary region and the turbine regions of the interface, while the steam generator region was not significantly different from the other two regions. In agreement with the hypothesis, participants could more accurately estimate component values for the primary region due to both its significant role in the simulation, and the additional attention it receives, as will be described in the subsequent section on attention analysis. Furthermore, the primary system has less reciprocal relationships, and therefore, it is a less complicated region of the interface than the turbine. The turbine is the most complicated aspect of the system since its state is governed by the combination of the configuration of all the components within the primary and steam generator regions. No significant interaction between set and region was found, $F(4,92) = 1.12$, $p = 0.35$.

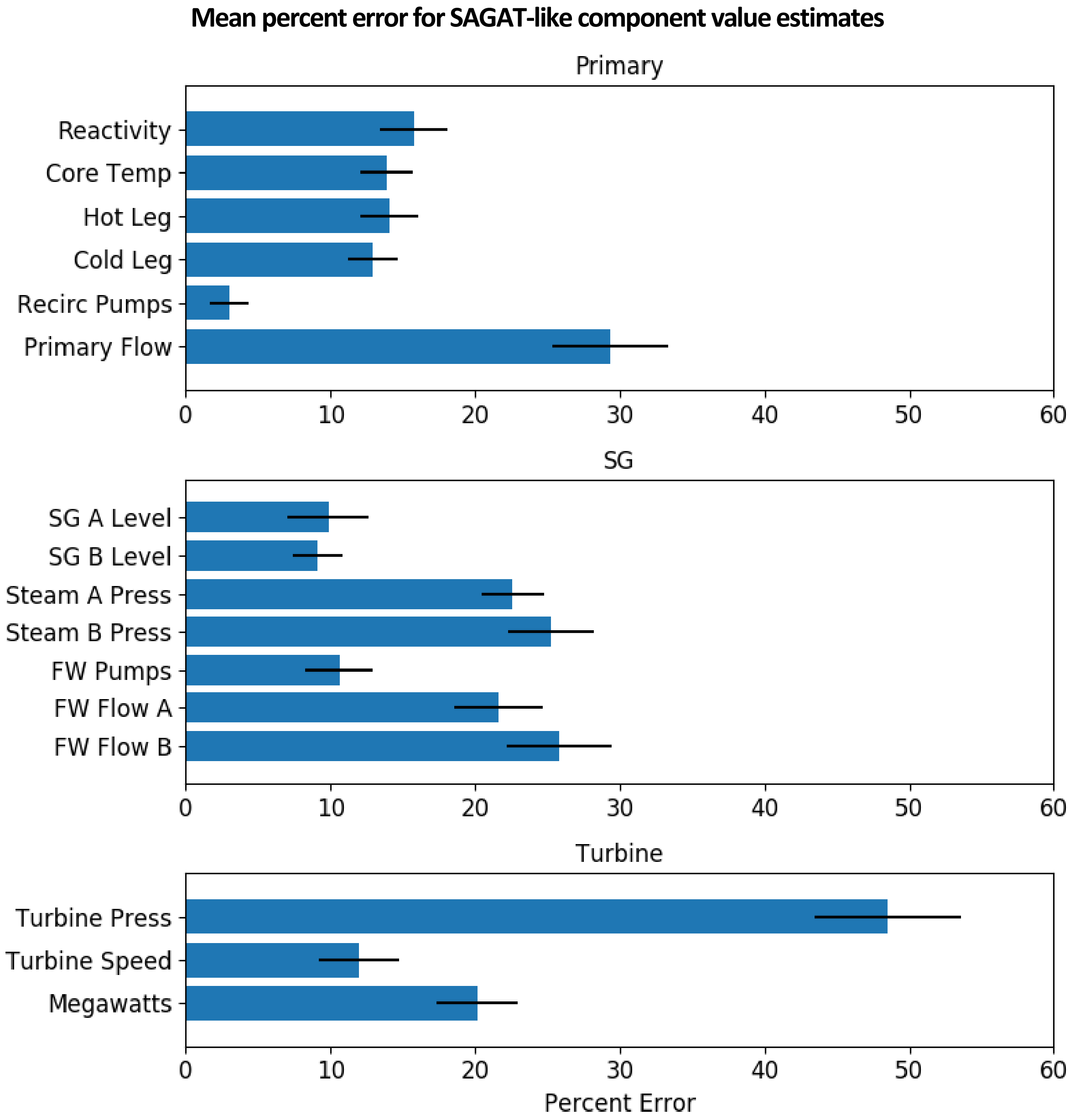


Figure 6.11 Component value estimates provided for the SAGAT-like SA probes were compared to the actual value from the simulation to yield a percent error. The mean percent error from SAGAT-like SA probes for each component in each of the three regions. Error bars indicate absolute error for each mean.

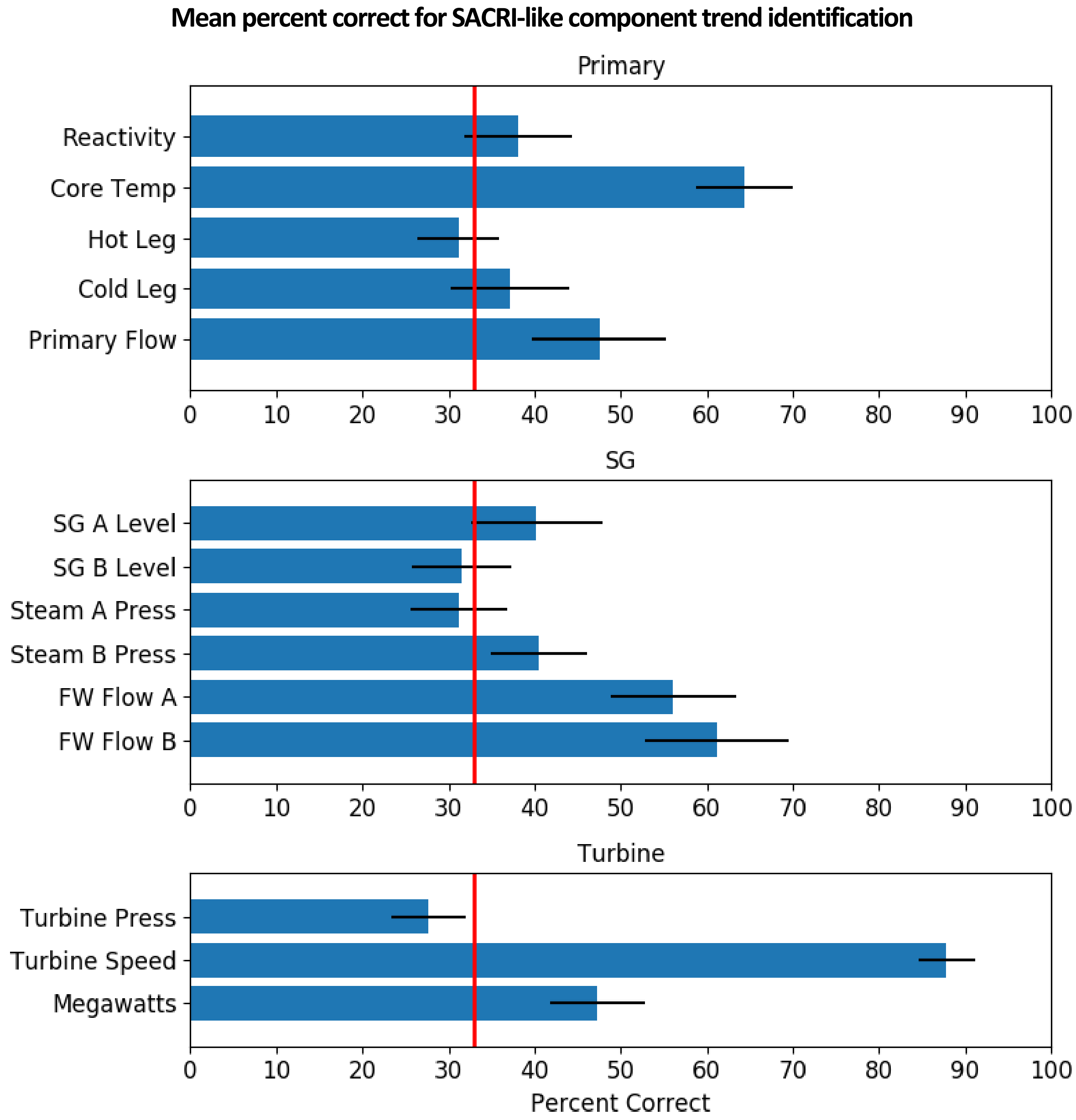


Figure 6.12 Trend identification responses for each SACRI-like SA probe for each component in the three regions were scored based on the actual trend of the component from the simulator log to yield the mean percent correct. The red line indicates chance response accuracy percentages. Error bars indicate absolute error for each mean.

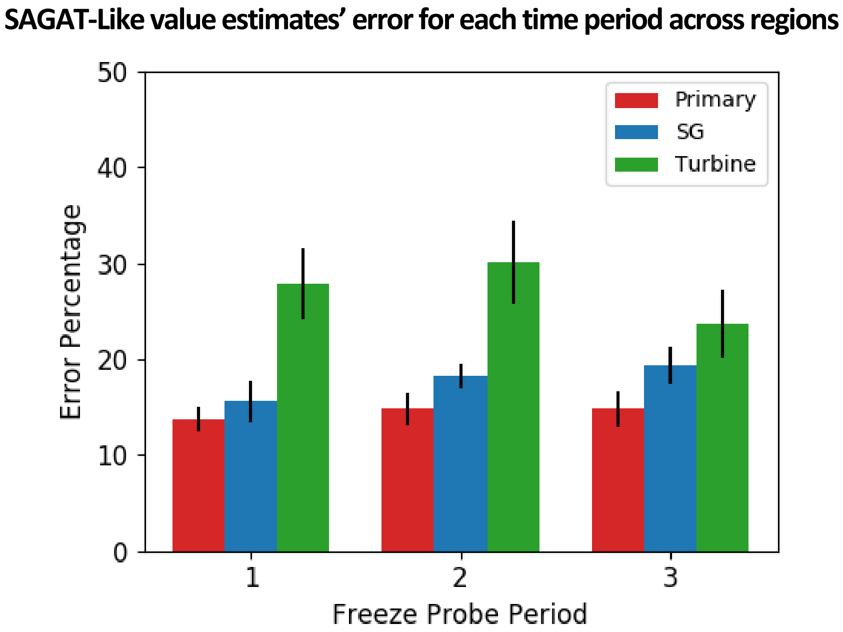


Figure 6.13 SA component value estimates' error. The percentage reflects the amount of error in the component value estimate provided during each of the three freeze probe administrations, set 1, set 2, and set 3. Error bars indicate standard error for each mean.

SACRI-like SA. To analyze freeze probe response trend identification accuracy, a 3 (region) by 3 (period) repeated measures ANOVA was performed. Again, it was hypothesized that the first administration, during startup, would demonstrate the best SA and the other two administrations would be lower. Furthermore, since the primary region is the least complicated region due to less reciprocal relationships with other components, it would demonstrate the best SA, while the turbine would demonstrate the poorest. A significant main effect for period was found, *F*(2,46) = 7.91, *p* = 0.00; however, the direction of the effect was contrary to the hypothesis. A post hoc Fischer's LSD test showed participants demonstrated significantly higher SA, as defined by higher accuracy in trend identification, on the third administration than the first administration, as can be seen in Figure 6.14. The first period contained a smaller collection of fluctuating components since the turbine was not in operation, yet the third period demonstrated more accuracy and involved fluctuating turbine region components. Most participants were able to recover from the fault and achieve a steady-state during the third period, which results in the entire system of components being fairly stable. As a result, participants were able to identify the trends quite accurately during this third period, while accuracy scores exceeded those of the first period. It is possible that participants failed to understand the components remained unchanged when the turbine is not in operation. Or simply, the turbine region component trends were never considered and participants did not even attempt to reason what their trend may be before guessing a trend during the first period, which was subsequently reflected by poorer accuracy scores. Participants were more aware of the turbine region components when they were in operation since they interacted with them, and since these components were stable while the plant was online in the third period, participants could more accurately report their trends.

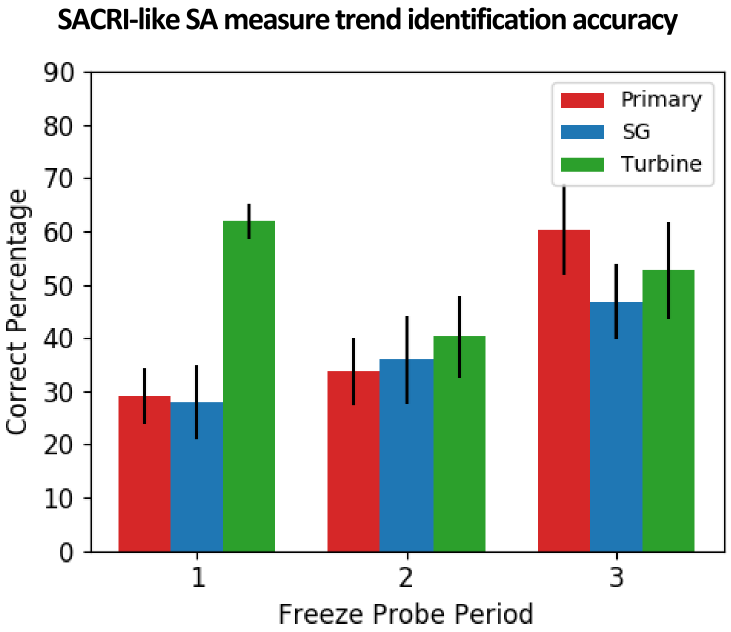


Figure 6.14 SA component trend identification accuracy. The percentage reflects the percentage of correct trend identifications. Participants report if the component was increasing, decreasing, or remaining unchanged. Error bars indicate standard error for each mean.

The greatest accuracy for the final probe administration is an interesting result because typically freeze probes are administered on larger time-scales than those found in the microworld. In particular, the SACRI is designed for full-scope simulations with time-frames spanning from half an hour to several hours (Hogg, 1995). Furthermore, the rate of change of parameters within a full-scope simulation is typically much slower than what was found in the microworld. The significant differences between freeze probe time periods provides supporting evidence that the SACRI probe style of question entailing making trend identification judgments is effective in a smaller scope and time-scale simulations since it was able to generate significant differences between the three time periods SA was assessed within the trials. This demonstrates the sensitivity of the SACRI probe style questions to shorter time periods within an eight-minute trial. Furthermore, the enhanced accuracy exhibited in the third period is likely due to the participant being able to achieve a more stable system, and therefore, make more accurate determinations about trends. During the first two periods, the system is in greater flux, which makes it more challenging to determine the correct trends and could account for differences observed for accuracy at the different time-points in which the probes were administered. No significant main effect for region was found, *F*(2,46) = 2.29, *p* = 0.11. A significant interaction between period and region was found, *F*(4,92) = 4.07, *p* = 0.00. The interaction results from significantly higher accuracy during the first period for the turbine region and significantly higher accuracy during the third period for the primary region probes. During the first period, the turbine is offline, and therefore, in a stable and unchanged state. As a result, participants potentially could easily identify the trend as unchanged, assuming they realized the turbine was offline, which resulted in the high-levels of accuracy for turbine region probes during the first period. During the last period, the system is in a stable-state and participants are more easily able to identify primary region trends that are fluctuating less than during the earlier periods in which the primary region is in flux for startup and syncing to the grid.

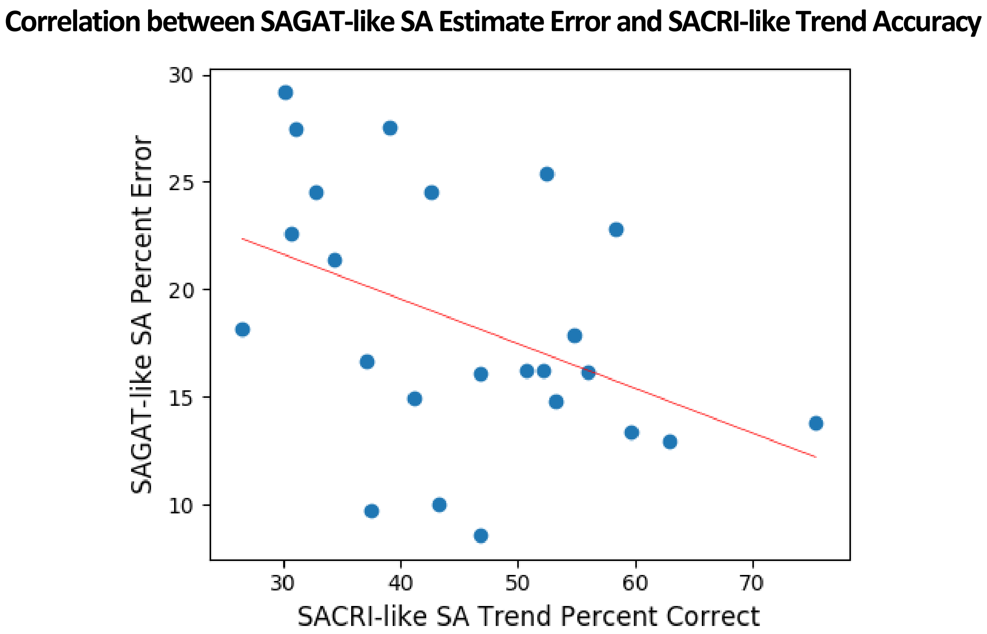


Figure 6.15 The SAGAT-like SA estimate error and SACRI-like trend accuracy were significantly and negatively correlated with each other.

The overall SAGAT-like estimate error and SACRI-like trend identification accuracy were found to significantly and negatively correlated with each other, r(24) = -0.43, 0.02. In general, undergraduate students did not perform as well on either the SAGAT-like or SACRI-like freeze probe responses as predicted. In particular, the SACRI-like responses were characterized by a number of participants responding at chance levels, which corresponds to 33% correct as can be seen in Figure 6.15. The poor performance of this measure contrasts with the findings of Hogg et al. and their studies concerning the development of the SACRI measure. Hogg et al. reported SACRI trend identification accuracies ranging between 60 and 100% (1995), while the range of responses for the SACRI measure administered to undergraduate students ranged between 30 and 75%. The results reported by Hogg et al. pertain to licensed operators performing scenarios on systems they were familiar with, while this study examined undergraduate students with no process control experience interacting with a system they were unfamiliar with. The challenge of working with the new system and the lack of experience can both contribute towards the poor performance of participants on these measures. The overall poor performance of the measures may account for the lack of significant relationships between the SA measures and the performance and attention measures, which will be discussed in the next section.

Situation Awareness and Performance Better SA was predicted to correlate with performance as measured by revenue. SA, as measured by both component error estimates and trend identification accuracy, were collapsed across periods to capture overall SA within each region over the course of each trial. Component error estimates and trend identification accuracy within each region was correlated with total revenue earned. No significant correlation was observed between component value error estimates in any of the regions and revenue (see Table 6.2). There was also no significant correlation observed between trend identification accuracy in any of the regions and performance. This lack of correlation between SA and performance was not entirely unexpected since others have also failed to find a significant correlation (O'Brien & O'Hare, 2007). O'Brien and O'Hare found that the SAGAT scores assessing SA at Level I did not correlate with performance in their study on aviation. Additionally, the lack of correlation can be due to the overall poor performance participants demonstrated with their responses to the SACRI-like measure, though there are additional reasons that can partially account for the apparent lack of a correlation. First, SA is necessary, but not sufficient, for good performance (Ensley, 1995a;

Ensley, 1995b). It is also possible that the performance gap resulting from the inability to make effective use of SA assumes that good SA at each level has been achieved. Again, Ensley's SA model has three levels: knowledge of the system parameters, integration of the system parameters into a cohesive representation, and the ability to predict the future system state (1995a). The measures of SA in this experiment do not accurately reflect the third level of Ensley's SA model, and therefore, it is possible that the participants' ability to predict actions and their effect on the system account for the performance differences.

Table 6.2 Correlations between SA value estimates' error and trend identification accuracy and performance measured by revenue. There were no significant correlations.

Type	Region	*r*	*p*
Component Value Estimates' Error			
	Primary	-0.19	0.38
	SG	-0.05	0.80
	Turbine	-0.07	0.75
Trend Identification Accuracy			
	Primary	0.19	0.38
	SG	-0.06	0.79
	Turbine	-0.16	0.46

The participant must also be able to take advantage of SA and translate that knowledge via physical actions into an appropriate control strategy for the system. The breakdown between SA and performance can result from the inability to make effective use of the SA. Indeed, the performance results show that participants with better revenue also demonstrated higher reactivity. This higher reactivity represents a different control strategy, which involves a more challenging dynamic system due to additional heat within the system that must be monitored and controlled. Though this strategy is more challenging, it also generates more steam and allows participants to produce more power, which leads to more revenue. Therefore, the differences in performance may be attributed to the differences in the ability to manipulate the controls under time-pressure to achieve high-reactivity and generate more revenue, rather than the SA levels. The attribution from better performance to a more sophisticated control strategy does not account for differences in SA and may explain the apparent lack of a correlation between SA and performance. Another possibility is that the revenue measure is not sufficiently sensitive to reflect changes in SA. To address this issue, the most sensitive performance measures—average reactivity and steam generator level error—were correlated with SA response accuracy to the probes pertaining to the primary and steam generator regions. No significant correlation was found between average reactivity and trend, *r*(24) = 0.03, *p* = 0.87, or error, *r*(24) = -0.01, *p* = 0.95, response accuracy to probes pertaining to the primary region. No significant correlation was found between steam generator level error and trend accuracy for probe questions concerning the turbine region, *r*(24) = -0.33, *p* = 0.11, but a significant correlation was found between steam generator level error and component value estimate error for probe questions concerning the turbine region, *r*(24) = 0.58, *p* = 0.00 (see Figure 6.16) Therefore, this more specific measure of performance does appear to reflect changes in SA.

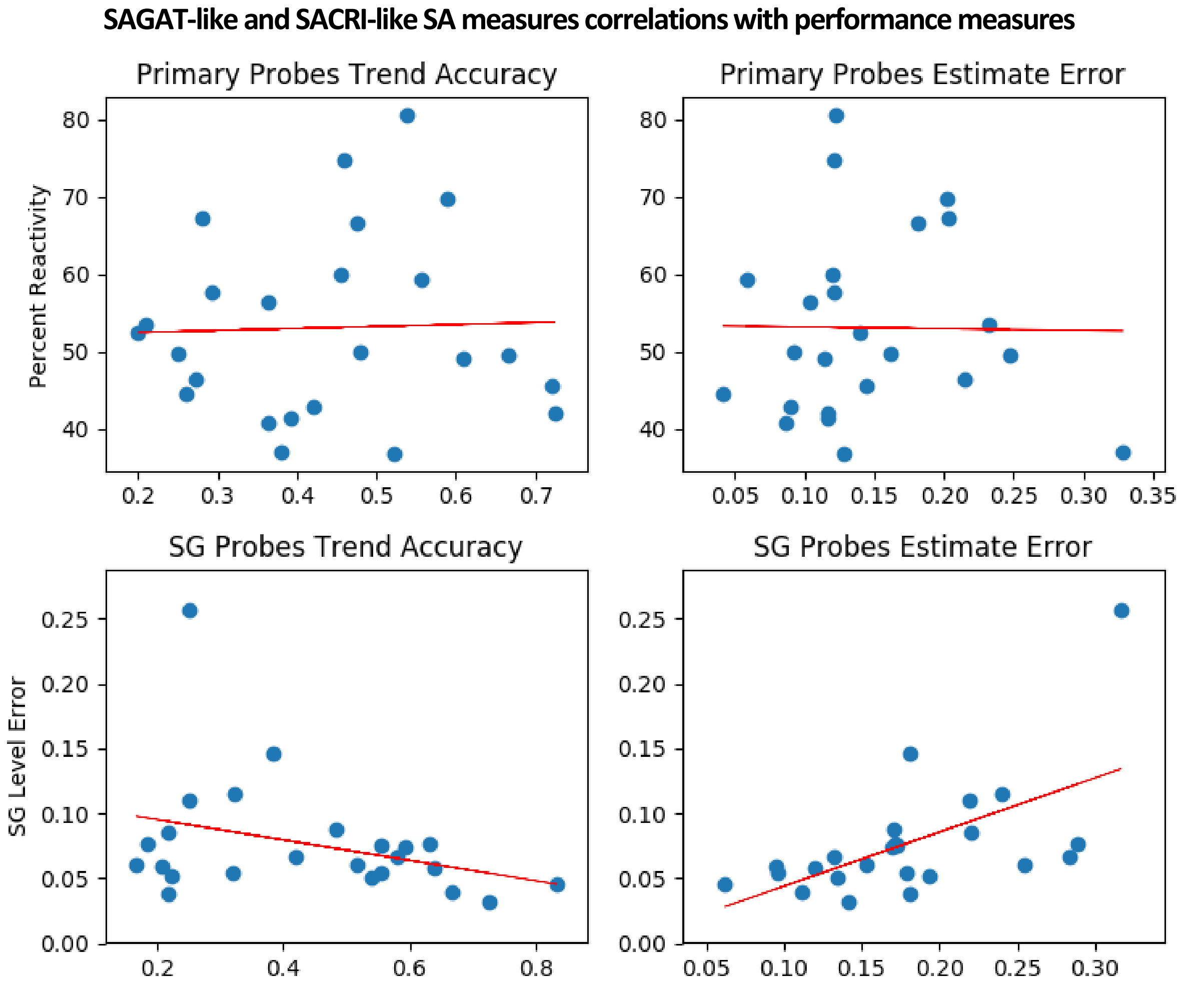


Figure 6.16 The two top panels depict the correlations between the two SA measures and the participant performance as measured by percent reactivity. The two bottom panels depict the correlations between the two SA measures and participant performance as measured by the error in maintaining the steam generator levels at the optimal 50% level.

Attention Analysis. The patterns of attention were examined to compare the attention-acknowledgement measure to the established eye tracking measure and to characterize the patterns of attention associated with the task for each mode of operation.

Correspondence. The new marker-acknowledgement attention measure was first evaluated against the traditional eye tracking technique by calculating correspondence scores between region classifications of each marker-acknowledgement and corresponding eye tracking data. To calculate the correspondence, the time of each marker-acknowledgement and the region in which the marker was located were identified. Then the eye tracking data was processed to identify the same point-in-time to determine if a fixation was recorded within the same region of the interface (i.e. alarms, primary, steam generators, and turbine regions). The result of this correspondence analysis provided average proportional rates of correspondence for marker-acknowledgements and eye tracking fixations. A window of time-points surrounding the acknowledgement time were also examined. These time-points spanned four seconds on either side of the marker to form an eight-second window around the marker-acknowledgement in which the eye tracking fixation data was assessed for correspondence. The eight-second time-window was assessed at 250-millisecond-increments within a two-second window surrounding the attention-acknowledgement time-point and 500-millisecond-increments for the remaining and more distant time-points. The results of this correspondence analysis mapping each acknowledgement with the corresponding fixation measured by eye tracking can be seen in Figure 6.17. The highest correspondence of 78% between fixations and acknowledgements was observed at approximately -1000 milliseconds prior to the marker-acknowledgement. The 78% correspondence value indicates that 78% of acknowledgements were accompanied by a fixation occurring within the same interface region near that time-point. The high-correspondence value provides evidence that the marker-acknowledgement and fixation data both sampled attention within the same interface region with the acknowledgement following the fixation after a brief time-lag. Since eye tracking fixations are an established measure, demonstrating both measures agree, provides evidence that the marker is able to measure attention similarly. It is also worth noting that the highest correspondence occurring at this earlier time-point suggests that participants are visually identifying the marker, performing the cognitive assessment of the target to verify it is in the target state, and then moving the mouse to select the marker. These activities take time, and therefore, the fixation precedes the physical acknowledgement of the marker. Furthermore, participants are possibly shifting attention towards the next screen element of interest as they execute the motor commands to select the marker, which can account for why a lower correspondence occurs during the actual marker-acknowledgement time.

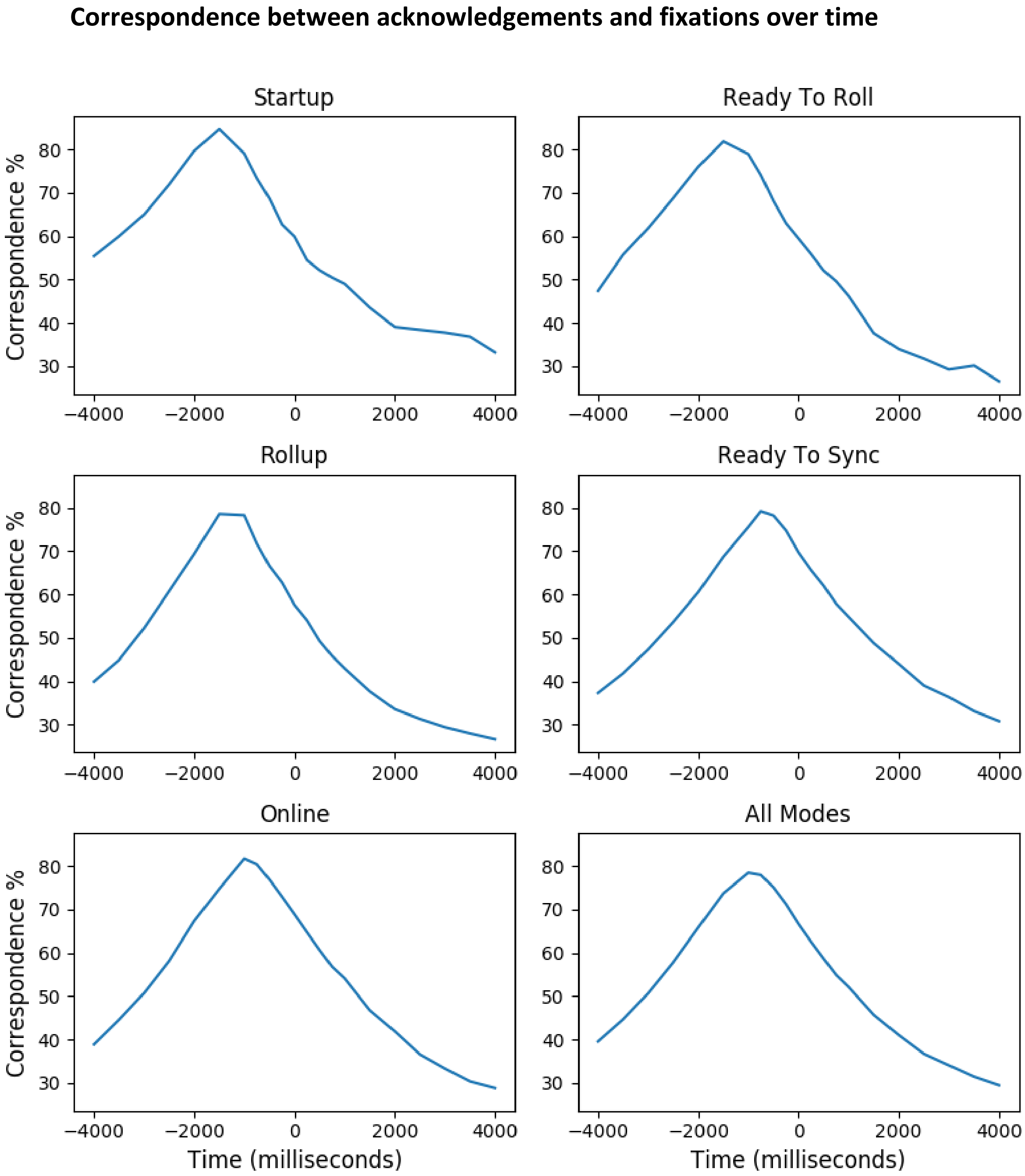


Figure 6.17 Correspondence analysis for each mode of operation across an eight-second time-window surrounding the acknowledgement time. Each plot shows the percentage of correspondence (i.e., percent of agreement) for the region classification of the acknowledgement and eye tracking fixations. Higher percentages indicate higher correspondence.

Acknowledgement and Fixation Correlations. Correspondence analysis addresses whether the eye tracking data demonstrates the same classification for each acknowledgement made by participants within the interface regions. The correspondence analysis does not assess instances in which the eye tracking measures a fixation in the absence of an accompanying acknowledgement. Eye tracking data is captured at a higher frequency, and therefore, it is necessary to compare the two measures over a larger time-span to determine if they are both measuring the same pattern of attention. To address the overall agreement in the classification of attention between the two measures, the proportions of attention within each region measured by acknowledgements and fixations were correlated with each. The proportion of fixations within each interface region was first calculated for each participant by dividing the number of fixations within each region by the total number of fixations performed across the entire trial, which includes all modes of operation. The proportion of acknowledgements for each participant was calculated using the same process. The proportions of fixations and acknowledgements for participants within each interface region were then correlated with each other to determine the overall agreement. The significant correlations between the proportions of attention measured by acknowledgements and fixations were observed between the primary, $r(26) = 0.59$, $p = 0.00$, and turbine, $r(26) = 0.57$, $p = 0.00$, regions of the interface. No significant correlation was observed between the two measures of attention for the proportion of attention allocated to the alarms $r(26) = 0.23$, $p = 0.26$ or steam generator regions $r(26) = 0.19$, $p = 0.36$. A graphical representation of the correlations between the two measures can be seen in

Figure 6.18. The lack of a correlation between the two measures for the alarm region was expected. The alarms are binary and transition between states throughout the trial. This state-transition can briefly capture attention; however, participants quickly return to their current locus of attention without performing an acknowledgement. Due to the differences in the sampling rate between the two attention measures, participants do not necessarily perform an acknowledgement when they briefly fixate within the alarm region, and therefore, a correlation was not found. The lack of a correlation between the two measures within the steam generator region is less clear. One likely explanation is the steam generators' central location. As participants transition between the primary and turbine regions, they fixate briefly upon the steam generators in the middle of the interface, but they do not take the extra effort to perform any acknowledgements. Instead, once the locus of attention lands within the primary and turbine regions after the transition, they spend more time focusing within these regions, and therefore, the opportunity to perform acknowledgements is greater.

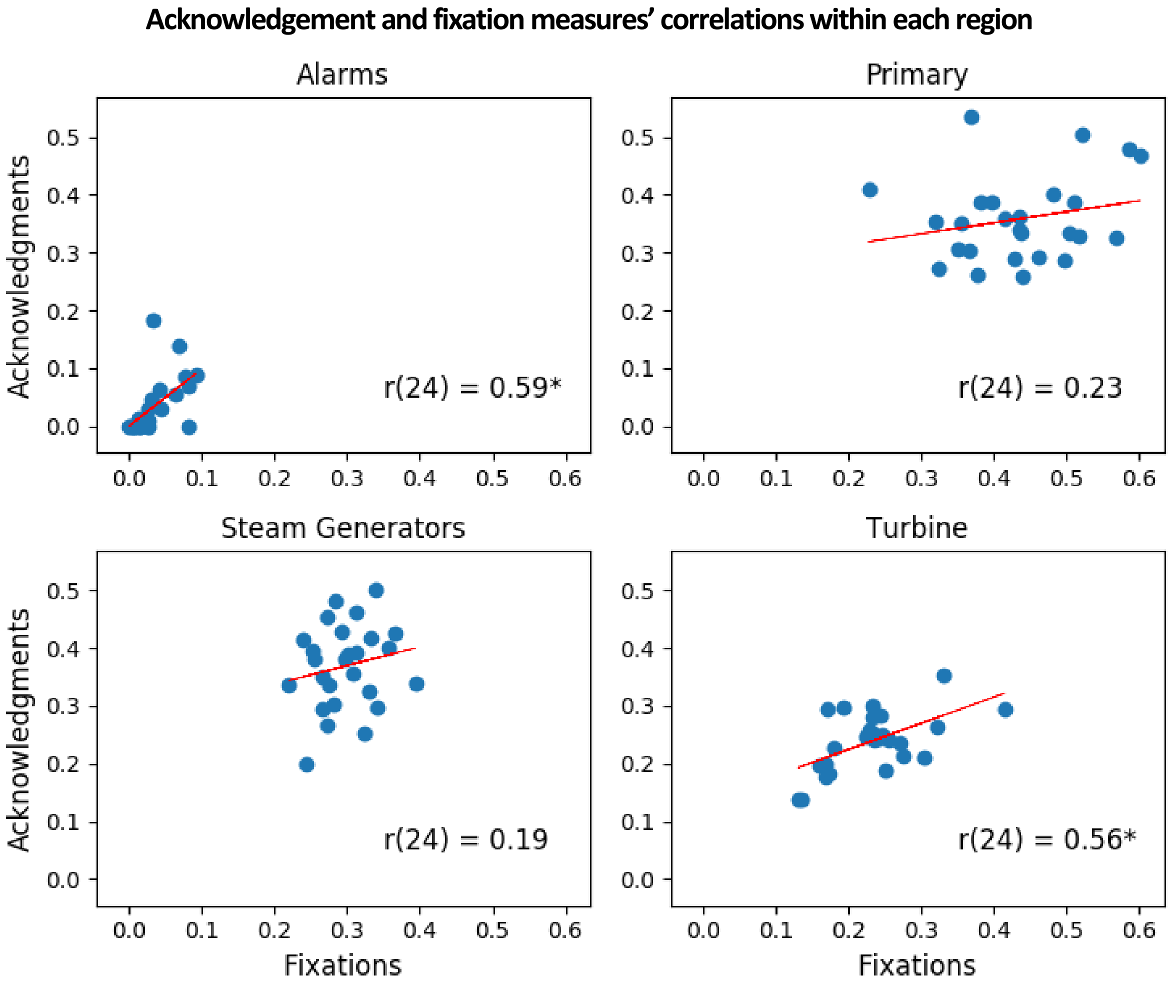


Figure 6.18 Correlations between acknowledgements and fixations in each of the four interface regions collapsed across all modes and trials. *p < 0.05

Attention within Modes. To understand the patterns of attention associated with specific tasks, the patterns of attention captured by the two measures were examined within each mode of operation. Unfortunately, the two short-duration modes—ready to roll (*M* = 16.2 s, *SD* = 8.0 s) and ready to sync roll (*M* = 8.1 s, *SD* = 6.2 s)—suffered from a lack of reliable data for the attention-acknowledgement measure. Many participants did not perform an acknowledgement during these short modes, as can be seen in Figure 6.19. Due to this lack of acknowledgement data, the ready to roll and ready to sync modes of operation were excluded from subsequent analysis. The lack of acknowledgements performed during these short-duration modes demonstrates one of the limitations of the acknowledgement measure. The acknowledgement measure has a low-temporal resolution, since it requires cognitive processing and deliberate responses. Therefore, the measure is not appropriate for short-duration tasks. The eye tracking temporal resolution is much higher at 17 milliseconds for the Tobii X2 eye tracking unit used in this research (Tobii, 2016). Though only one eye tracking device was used in this research, this temporal resolution is common among commercially available eye-trackers and representative of the technology's capabilities. There were no instances in which the eye tracking failed to capture any samples during the short-duration modes. As a result, eye tracking should be used during short-duration tasks as opposed to the acknowledgement measure with its lower temporal resolution.

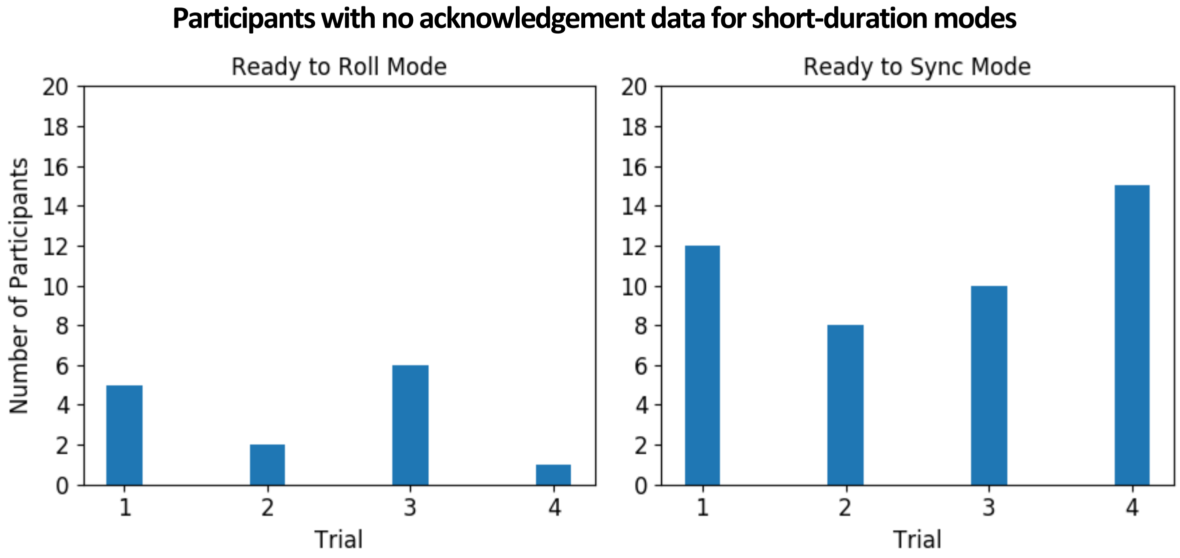


Figure 6.19 The ready to roll and ready to sync modes were characterized by a general lack of acknowledgement data due to the short-duration of these two modes of operation. This

figure shows the number of participants without a single acknowledgement performed during the ready to roll and ready to sync modes.

The patterns of attention defined as the proportion of attention allocated to each of the four regions of the interface were examined during each of the activities associated with the startup, rollup, and online modes of operation within each of the four experimental trials. Though the participants did demonstrate improved performance on the later trials, it was necessary to collapse the attention-measures across the trial to ensure sufficient observations were made for each cell of the statistical model. The patterns of attention were predicted to vary based on the current process control activity in accordance with the SEEV model and the predictions it makes in where attention should be allocated based on the interface characteristics and task demands. During the startup mode of operation, a higher proportion of attention was predicted to be allocated to the primary region, since this region of the interface contains the indicators and controls necessary to support controlling reactivity and heating the reactor coolant. During the rollup and online modes of operation, the distribution of attention was predicted to be more even across the primary, steam generator, and turbine regions. The alarm region was predicted to receive less attention than the other three during all modes of operation, since it provides redundant information and is of little extra value for monitoring and controlling the simulation. The attention-allocated-proportions calculated for each region are inherently dependent upon each other within each mode of operation. Analyzing differences in proportions of attention within regions for each mode in a single ANOVA violates the assumption of independent cases, and therefore, it was necessary to examine each region within modes independently in separate

ANOVAs to ensure the ANOVA assumption for independent cases was upheld in each analysis. Therefore, separate 2 (measurement type) by 3 (mode) repeated measures ANOVA were performed to analyze the proportions of attention within each region across the operation modes. The results of the ANOVA analyses can be seen in Table 6.3, while a graphical representation of the main effects and interactions can be seen in Figure 6.20. The alarm region exhibited a main effect for mode. A post hoc analysis using Fischer's LSD showed the online mode was significantly different from startup and rollup while startup and rollup were not significantly different from each other. Participants showed higher proportions of attention to the alarm region during the online mode due to alarms triggering when faults occurred or if the participants encountered any issues as they manipulated the simulation. These alarms occurred less frequently during the startup or rollup modes; therefore, less attention was drawn to the alarm region.

Table 6.3 Results from the ANOVA analyses of the two measurement types—acknowledgement and fixation proportions—across the three modes of operation. The * indicates p < 0.05.

	Source	df	df_{error}	*F*	*p*	η^2	β
Alarms	Type	1	25	3.42	0.08	0.02	0.88
	Mode	2	50	8.30	0.00*	0.09	0.96
	Type * Mode	2	50	0.04	0.97	0.00	0.05
Primary	Type	1	25	7.94	0.01*	0.06	1.00
	Mode	2	50	61.82	0.00*	0.86	1.00
	Type * Mode	2	50	3.71	0.03*	0.02	0.38
Steam Generators	Type	1	25	15.08	0.00*	0.15	1.00
	Mode	2	50	23.16	0.00*	0.22	1.00
	Type * Mode	2	50	2.12	0.13	0.13	0.23
Turbine	Type	1	25	0.00	0.99	0.00	0.05
	Mode	2	50	60.96	0.00*	0.63	1.00
	Type * Mode	2	50	5.79	0.01*	0.01	0.56

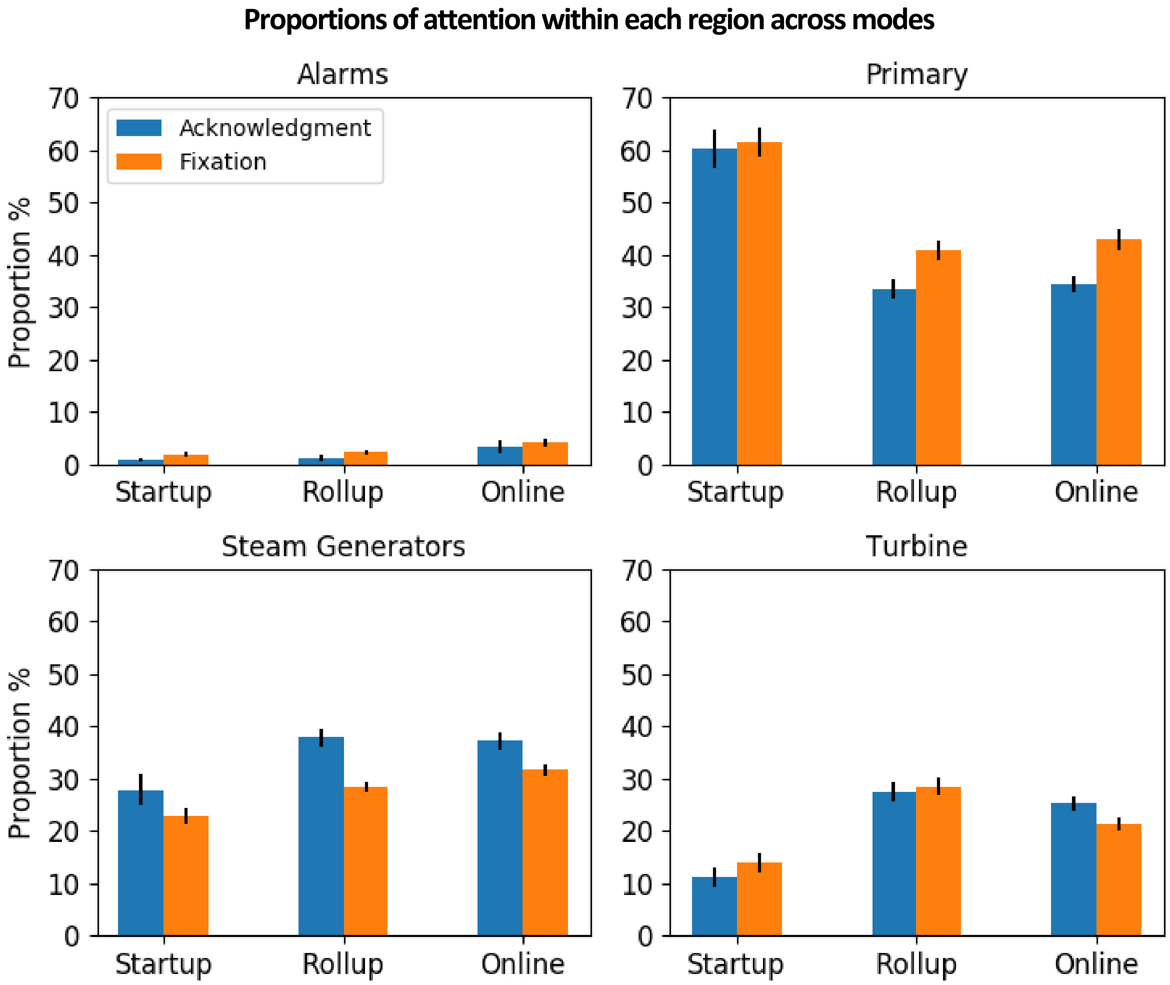


Figure 6.20 Proportions of attention by region for each mode of operation and showing the main effects and interactions for measurement-type and mode of operation. Error bars indicate standard error for each mean.

The primary region demonstrated significant main effects for measurement-type and mode with a significant interaction between the two factors. A post hoc analysis using Fischer’s LSD showed the interaction results from the two measurement types demonstrating significantly different proportions of attention within the rollup and online modes of operation, while both measures did not significantly differ in the proportion of attention measured in the startup mode of operation, which is where attention is mostly dedicated to the primary region with little reason to shift attention to the other regions since none of the components outside of the primary region are necessary for the task. Since the other modes do involve some shifts of attention, since tasks span across regions, the differences in the proportions measured by the acknowledgements and fixations diverge since fast shifts in attention will be captured by fixation data, while the ability to process a marker and physically manipulate it for the acknowledgement is a slower process and may not be captured.

The steam generators region showed a main effect for type and mode with no significant interaction. The acknowledgement measure consistently exhibited a higher proportion of attention than the fixation mode within the steam generator region across all three modes. Lastly, the turbine region showed a main effect for mode and a significant interaction between mode and type. A post hoc analysis using Fischer's LSD showed each mode with significantly different proportions of attention from the other two modes. As predicted, the turbine received the smallest proportion of attention during the startup mode and the greatest proportion of attention during the rollup mode, as would be expected due to the nature of the tasks during these modes. The turbine is not related to the task of raising control rods to begin producing steam during startup and received little attention. During rollup, the turbine is the primary component of interest, and therefore, received the greatest amount of attention when compared to any other modes.

The results within each region, as summarized previously in Table 6.3, demonstrate the differing effectiveness of the fixation and acknowledgement measures to assess attention. A significant effect for mode was found for each region, which provides evidence that these measures were sensitive to the different patterns of attention exhibited by participants because of the different activities they were performing within each mode. The main effects for type and interactions between mode and type reflect how the two measures diverge as they attempt to capture patterns of attention during different activities. Though there were significant effects for type in the primary and steam generator regions and significant interactions in the primary and turbine regions, the overall patterns of attention captured by measures remained the same across modes. For example, in the primary region, even though a main effect for type and an interaction between type and mode were found, both measures assessed the highest proportion of attention for the primary in the startup mode, and smaller but roughly equivalent proportions in the rollup and online modes. The relative patterns of attention are important because they reflect the acknowledgements as participants perform them, and therefore, the attention-acknowledgement measure was effective in assessing attention similarly to the established eye tracking-based fixation measure.

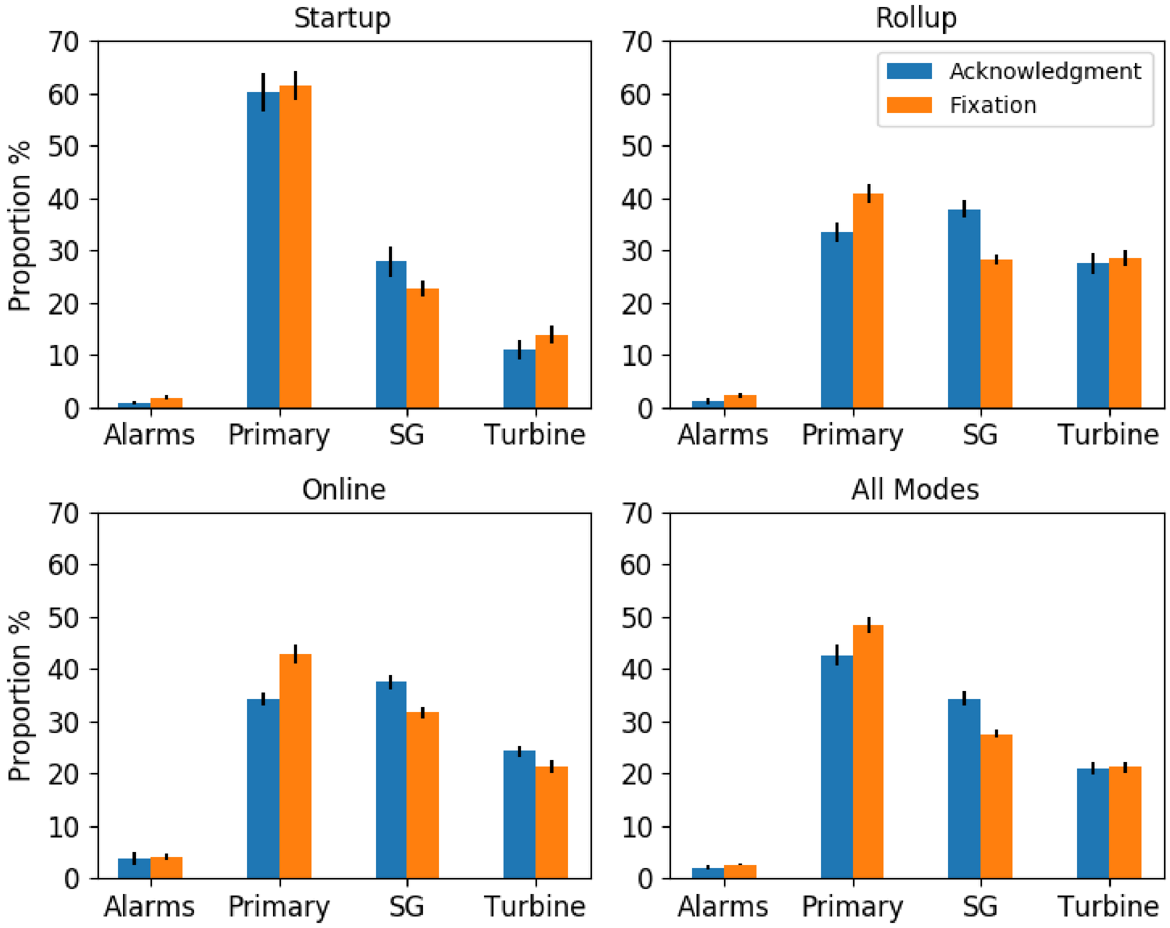


Figure 6.21 Profiles of the pattern of attention for each mode of operation and aggregated across all modes for a global representation of the profile of attention over the duration of the trial (lower right). Error bars indicate standard error for each mean.

Another useful way to examine the patterns of attention captured by the two measures is to visualize the proportions of attention by creating profiles for each mode of operation based on the analysis of the proportions of acknowledgements and fixations, which can be seen in Figure 6.21, the alarm region received little attention in all modes of operation. The alarms provide redundant information and were ignored by participants as expected. The startup mode of operation is characterized by a significantly larger amount of proportion allocated the primary region of the interface. Since the reactor core is located in the primary region and is the subject of the majority of monitoring and control activities, a greater proportion of attention directed to the primary region results. The steam generators and turbine, respectively, are less relevant during this mode of operation, which is also reflected in the lower proportion of attention allocated to the steam generators than the primary and the lowest proportion of attention dedicated to the turbine. The pattern of attention within the online mode of operation is more evenly distributed across the regions as predicted. The primary and steam generator regions exhibited the two highest proportions of attention during the online mode of operation. The turbine must only be monitored during the online mode to detect a trip event and otherwise does not provide any additional information. As a result, of the three relevant interface regions, it demonstrates the lowest proportion of attention.

Proportions of attention during each mode across regions

Attention within Recovery. The two patterns of attention measured by the acknowledgements and eye tracking fixations was also examined during the scripted fault recovery periods. Like the short-duration modes, the reactor recovery time reactor fault (*M* = 27.79, *SD* = 15.18) also suffered from missing observations due to participants lack of performing acknowledgements. The turbine fault recovery time was sufficiently long so that all but one participant performed multiple acknowledgements. As a result, only the turbine recovery period was analyzed for patterns of attention. The recovery period in the turbine fault condition was predicted to exhibit higher proportions of attention allocated to the turbine region, since identification of the fault and recovery actions occur within the turbine region. The data was collapsed across trials to prevent missing cells, and therefore, the recovery period was not examined across trials. A 2 (measurement type) by 4 (region) repeated measures ANOVA was performed to examine the proportions of attention during the turbine recovery period. No significant main effect for type was found, *F*(1,23) = 0.39, *p* = 0.54. A significant main effect for region was found, *F*(3,69) = 74.00, *p* = 0.00. Also, a significant interaction between region and type was found, *F*(3,69) = 13.65, *p* = 0.00. A Fischer's LSD post hoc analysis of the type of measure showed the means for the acknowledgement and fixations within the primary and steam generator regions were significantly different from each other. Additionally, Fischer's LSD post hoc analysis of the region factor showed the means for the alarms and primary regions were significantly different from all other region means. The steam generator and turbine regions were not significantly different from each other. The significant interaction and significant differences in the proportions of attention assessed within the alarms and primary regions demonstrate the divergence between the two measures of attention during a fast-paced task requiring many actions. Indeed, since the acknowledgement measure requires actions to capture attention, it likely did not accurately capture attention in this type of situation as effectively as the fixation measure. This type of situation represents one in which the endowment measure performs poorly and would not be recommended. The acknowledgement measure instead functions better over long time-periods in which the primary task is monitoring and few actions are required.

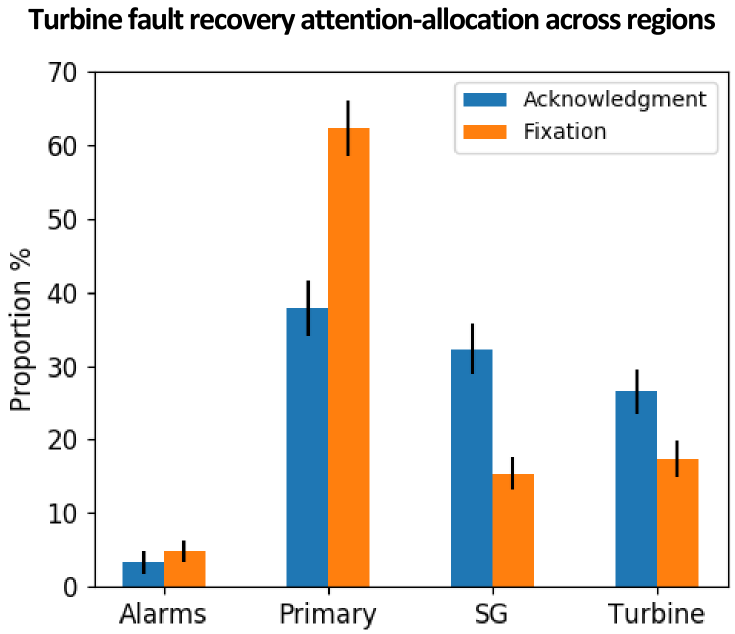


Figure 6.22 The proportion of attention allocated to each region of the interface during the scripted turbine fault recovery period. Error bars indicate standard error for each mean.

Figure 6.22 shows the pattern of attention during the turbine recovery period. Counter to the hypothesis, the turbine region did not receive more attention during the recovery period. In fact, the primary region underwent an increase in comparison to the online mode of operation for the fixation measure of attention, as shown previously in Figure 6.21. This appears counterintuitive; however, the turbine fault requires participants to adjust reactivity to compensate for the turbine no longer receiving the steam once the fault occurs and closes the valves leading into the turbine. As a result, the participants must reduce reactivity while the turbine is ramping up to the 1800 rpm speed required for syncing and resuming the online mode of operation. It is possible for participants to quickly resume normal operations if they detect and re-latch the turbine at the onset of the fault, but the length of the turbine recovery period in seconds (*M* = 58.06, *SD* = 28.37) suggests this was not the typical recovery situation experienced by participants, since the average recovery time is nearly 60 seconds. Had participants followed this proposed quick re-latch strategy, the average recovery time would be much shorter. This also demonstrates a limitation of the acknowledgement measure since it failed to detect the shift towards the primary region unlike the fixation measure. Thus, during these more rapid- and short-duration tasks, the fixation measure demonstrates superior assessment of attention than the acknowledgement measure.

Attention and Temporal Resolution. As a secondary task, the acknowledgement measure of attention requires cognitive processing and a physical response via a mouse click to assess attention. Conversely, the eye tracking measure of attention is a passive measure that requires not cognitive effort or a physical response. Furthermore, the eye tracking measure of attention has a high-sampling rate of 17 milliseconds. Together these aspects of the two measures suggest the acknowledgement measure should have a poorer temporal resolution than the traditional eye tracking fixation-based measure of attention. Indeed, the poorer temporal resolution was found, as can be seen in Figure 6.23.

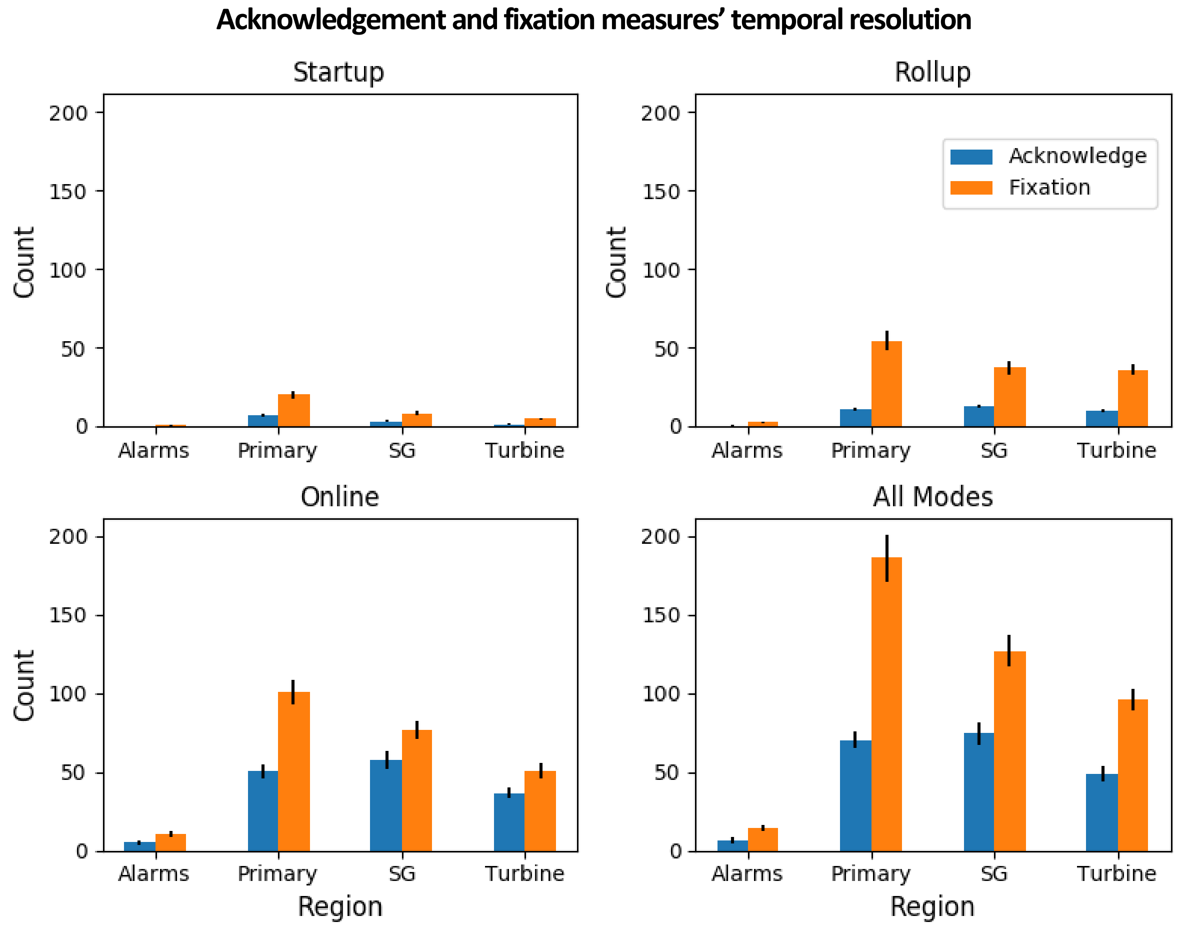


Figure 6.23 Comparison of acknowledgement and fixation measurement resolution in terms of the number of samples each captured during the three analyzed modes of operation and all modes (depicted in the lower right pane). Error bars indicate standard error for each mean.

As predicted, the acknowledgement measure demonstrated a lower temporal resolution (*M* = 24.97, *SD* = 10.05 acknowledgements per minute) than the eye tracking measure (*M* = 52.95, *SD* = 15.98 fixations per minute). The difference in the counts per minute between the acknowledgement and fixations shows that the eye-tracker is capable of capturing attention at twice the temporal resolution of the acknowledgement measure. To more thoroughly compare the resolution between the acknowledgement and eye tracking fixation measures of attention, the number of samples captured by each measure in the three modes of operation were compared across regions of the interface. A 2 (measurement type) by 3 (mode) by 3 (region) repeated measures ANOVA was performed. As expected, the acknowledgement measure showed reliably lower counts compared to the fixation measure, *F*(1,25) = 56.22, *p* = 0.00. Additionally, significant main effects for mode, *F*(2,50) = 107.93, *p* = 0.00, and region, *F*(3,75) = 66.62, *p* = 0.00, were found. The difference in the counts within mode and region reflect the activities of participants throughout the trial rather than any differences of the two attention measures. However, significant interactions between type and mode, *F*(2,50) = 34.61, *p* = 0.00; type and region, *F*(3,75) = 26.33, *p* = 0.00; mode and region, *F*(6,150) = 40.67, *p* = 0.00; and type and mode and region, *F*(6,150) = 19.84, *p* = 0.00, provide evidence that the two measures have different capabilities to measure attention within regions and modes of operation. The three-way interaction between type and mode and region can be explained by several factors. The main source of the interaction stems from the acknowledgement and fixation count means not differing significantly in the alarm region of the interface, regardless of the mode of operation. The acknowledgement and fixation counts in the primary, steam generators, and turbine were significantly different in all modes of operation, but the magnitude of the differences shifted depending on the mode of operation. The acknowledgement and fixation counts in the primary region differed more during the rollup and online modes of operation than the ready to roll and ready to sync modes of operation. During longer modes, a far greater number of samples can be captured by the fixation measure, which leads to a disproportionally larger difference than what was found for shorter duration modes.

Attention and Performance. In line with the SEEV model, a more optimal pattern of attention was predicted to accompany better performance, since attending to more pertinent elements of the interface fosters better performance by enabling participants to acquire information efficiently and inform their actions in controlling the simulation. Therefore, it was predicted that participants with better performance would exhibit a different pattern of attention than participants with poorer performance. To address this hypothesis, the participants were ranked by performance as measured by revenue. The top-third and bottom-third of participants were compared based on the proportions of attention allocated to each of the four interface regions within each mode of operation. Qualitatively, there is no difference between the top and bottom performing groups (see Figure 6.24), which indicates that a more optimal pattern of attention was not evident for top-performing student participants. The analysis first examined the patterns of attention measured by fixations in separate 2 (performance group) by 3 (mode) repeated measures ANOVA for each of the four interface regions. A significant main effect for mode was found in all analyses (see Table 6.4), which was not surprising since this effect was also found in the prior analysis comparing acknowledgement and fixation measures of attention. More important for this analysis is the comparison between groups and the interaction between group and region.

No significant effect for group was found nor were any significant interactions between mode and group found.

Table 6.4 Results from the ANOVA analyses of the proportions of attention measured by fixations for the top- and bottom-third of performing participants. No main effect for group was observed, indicating both top and bottom performers exhibited similar patterns of attention. The * indicates p < 0.05.

	Source	df	df_{error}	*F*	*p*	η^2	β
Alarms	Mode	2	28	4.75	0.02*	0.25	0.75
	Group	1	14	2.80	0.12	0.17	0.34
	Mode * Group	2	28	0.33	0.72	0.02	0.10
Primary	Mode	2	28	47.69	0.00*	0.77	1.00
	Group	1	14	2.50	0.14	0.15	0.31
	Mode * Group	2	28	0.37	0.70	0.03	0.10
Steam Generators	Mode	2	28	20.92	0.00*	0.60	1.00
	Group	1	14	0.00	0.99	0.00	0.05
	Mode * Group	2	28	0.15	0.87	0.01	0.07
Turbine	Mode	2	28	44.05	0.00*	0.76	1.00
	Group	1	14	2.00	0.18	0.13	0.26
	Mode * Group	2	28	1.40	0.26	0.09	0.28

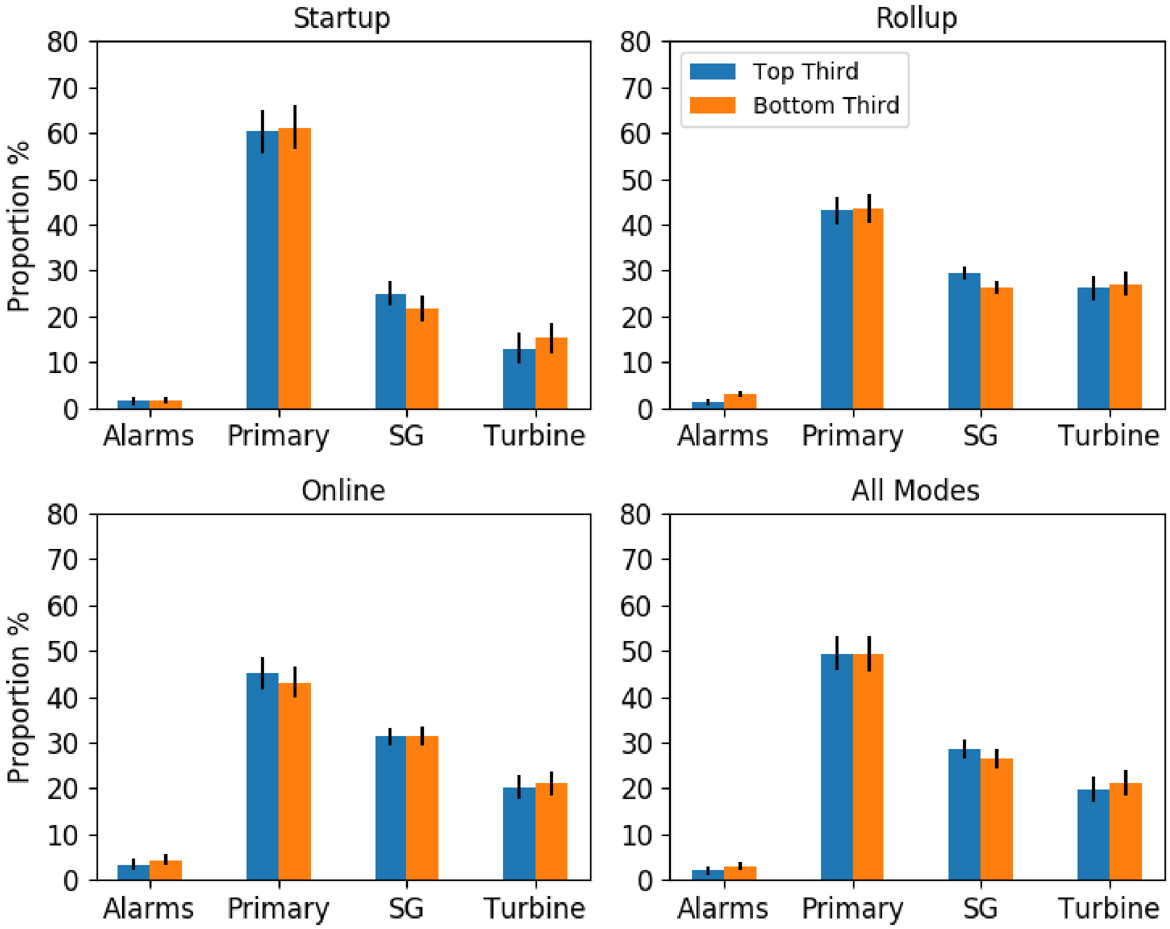


Figure 6.24 Comparison of the top- and bottom-performing student groups in terms of the pattern of attention measured by the fixation measure within each of the modes of operation.

Next, an additional analysis was performed to examine the patterns of attention assessed by the acknowledgement measure between the top- and bottom-performing students. The patterns of attention exhibited by the top- and bottom-performing groups of students can be seen in Figure 6.25.

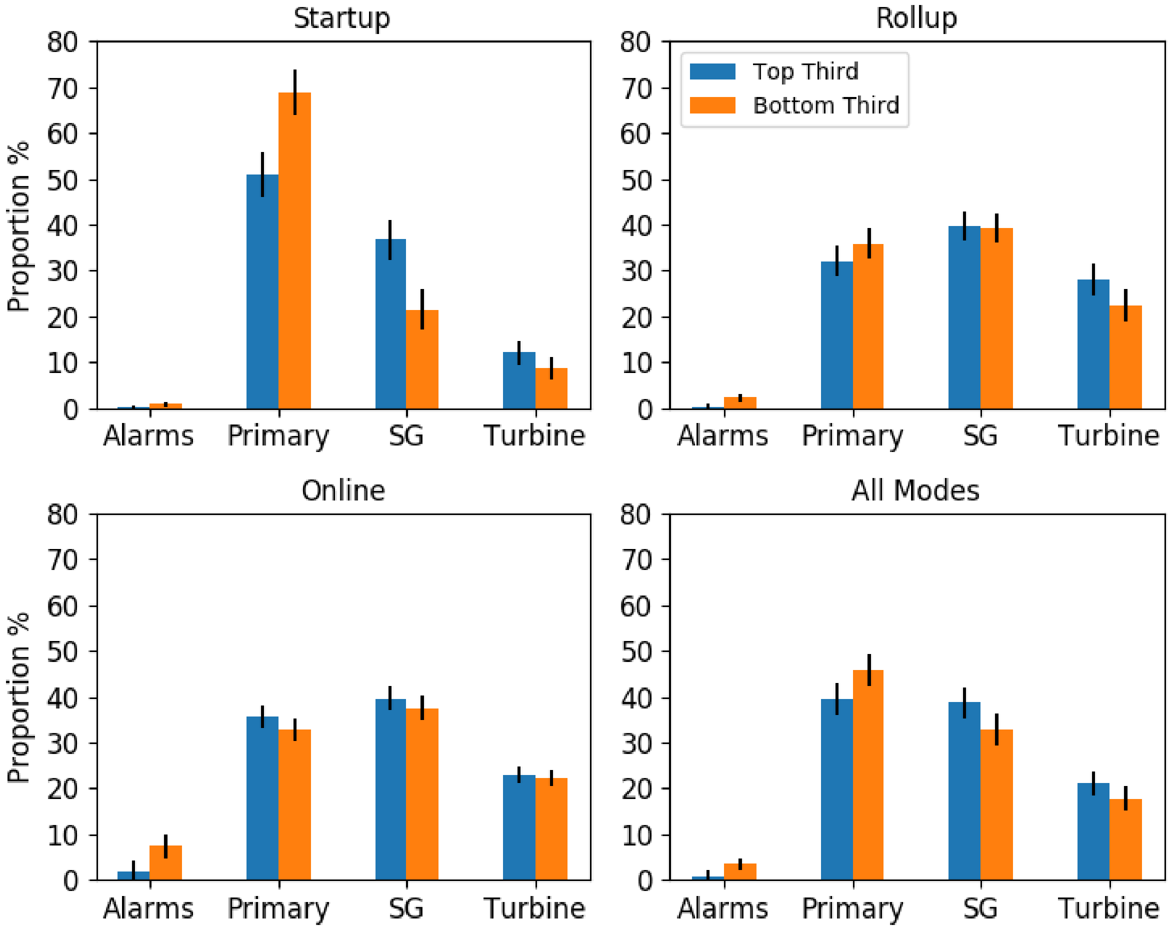


Figure 6.25 Comparison of the top- and bottom-performing student groups in terms of the patterns of attention measured by the acknowledgement measure within each of the modes of operation.

The patterns of attention measured by acknowledgements were analyzed in separate 2 (performance group) by 3 (mode) repeated measures ANOVA were for each of the four interface regions. The results of this analysis can be seen in Table 6.5.

Table 6.5 Results from the ANOVA analyses of the proportions of attention measured by fixations for the top- and bottom-third of performing participants. No main effect for group was observed indicating both top and bottom performers exhibited similar patterns of attention. The * indicates p < 0.05.

	Source	df	df_{error}	*F*	*p*	η^2	β
Alarms	Mode	2	28	4.77	0.02*	0.25	0.75
	Group	1	14	2.96	0.11	0.18	0.36
	Mode * Group	2	28	1.65	0.21	0.11	0.32
Primary	Mode	2	28	35.08	0.00*	0.72	1.00
	Group	1	14	3.80	0.07	0.21	0.44
	Mode * Group	2	28	4.42	0.21	0.24	0.71
Steam Generators	Mode	2	28	8.64	0.01*	0.38	0.95
	Group	1	14	2.47	0.14	0.15	0.31
	Mode * Group	2	28	4.28	0.02*	0.23	0.70
Turbine	Mode	2	28	40.27	0.00*	0.74	1.00
	Group	1	14	0.96	0.34	0.06	0.15
	Mode * Group	2	28	0.97	0.37	0.07	0.20

Again, a significant main effect for mode was found in all analyses, but this is not of much interest for comparing the patterns of attention between top- and bottom-performing groups. A single significant effect for group was not found in any of the analysis. A significant interaction between group and mode was found for the steam generator region analysis. A Fischer's LSD showed the interaction resulted from a significantly higher proportion of attention exhibited by the top performers during the startup mode, while the other two modes did not show a significant difference. This was the only significant interaction, since no main effect for group was found in any of the analyses. Together, the analyses for both the acknowledgement and fixation measures of attention demonstrate that there is little difference in patterns of attention exhibited by the top- and bottom-performance groups.

In contrast to the predicted hypothesis, there were no significant differences found for the proportions of attention allocated to each interface region between the top- and bottom-performing participant groups. This lack of difference indicates that differences in performance cannot be attributed to differences in the patterns of performance, since both the top- and bottom-performing individuals exhibit similar patterns. Since performance as measured by revenue is a somewhat blunt measure, the performance measure may not have had sufficient sensitivity to differentiate participants based on the efficiency of their patterns of attention-allocation. It is also possible that attention is not the limiting factor for performance in the microworld simulation. Instead, performance may be accounted for by other factors, such as SA and the participant's ability, translate that SA into successful manipulations of the simulation.

Attention and Situation Awareness. Using the SEEV model as a rationale for attention-allocation as a function of interface characteristics and task demands, participants with better SA were expected to demonstrate a more optimal pattern of attention than participants with poorer SA. Participants were ranked based on their SA scores and then the top- and bottom-thirds of participants were identified. The top- and bottom-third were then compared in terms of the patterns of attention they exhibited within each of the interface regions across all modes of operation. The first set of analyses examined the patterns of attention measured by fixations. Overall, the patterns of attention between the top and bottom SA student groups were similar, as can be seen in Figure 6.26.

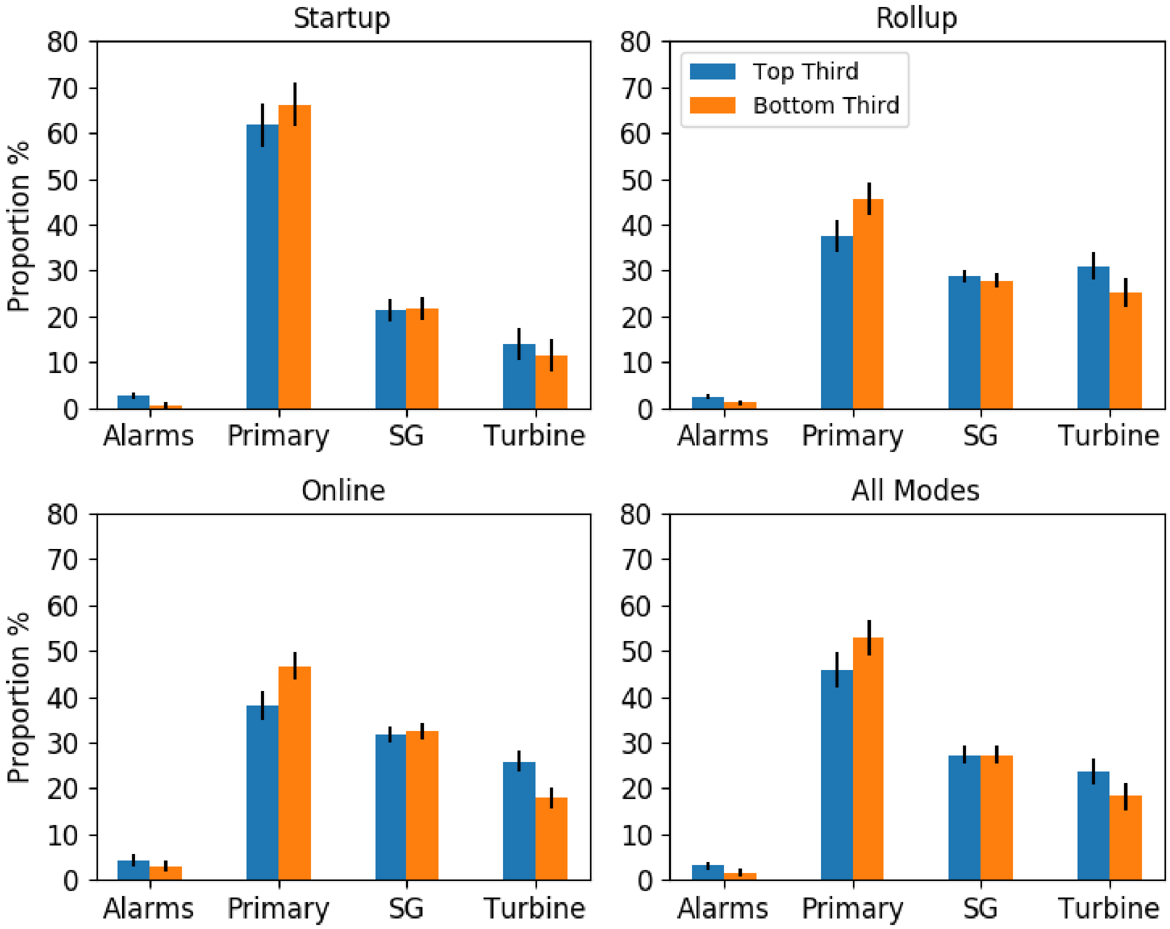


Figure 6.26 Comparison of the top and bottom SA student groups in terms of the patterns of attention measured by the acknowledgement measure within each of the modes of operation.

A 2 (performance group) by 3 (mode) repeated measures ANOVA was performed for each of the four interface regions. The results of these analyses can be seen in Table 6.6. As with performance, there is no significant difference between the top and bottom SA groups, indicating that both groups exhibited similar patterns of attention.

Table 6.6 Results from the ANOVA analyses of the proportions of attention measured by fixations for the top- and bottom-third of participants based on SA scores. No main effect for group was observed indicating both top and bottom performers exhibited similar patterns of attention. The * indicates p < 0.05.

	Source	df	df_{error}	*F*	*p*	η^2	β
Alarms	Mode	2	28	4.75	0.02*	0.25	0.75
	Group	1	14	2.80	0.12	0.17	0.34
	Mode * Group	2	28	0.33	0.72	0.02	0.10
Primary	Mode	2	28	47.69	0.00*	0.77	1.00
	Group	1	14	2.50	0.14	0.15	0.31
	Mode * Group	2	28	0.37	0.70	0.03	0.10
Steam Generators	Mode	2	28	20.92	0.00*	0.60	1.00
	Group	1	14	0.00	0.99	0.00	0.05
	Mode * Group	2	28	0.14	0.87	0.01	0.07
Turbine	Mode	2	28	44.05	0.00*	0.76	1.00
	Group	1	14	2.00	0.18	0.12	0.26
	Mode * Group	2	28	1.40	0.26	0.09	0.28

Next, the analysis examined the patterns of attention between the top- and bottom-performing students. The patterns of attention exhibited by the top- and bottom-performing groups of students can be seen in Figure 6.27. As with the previous analysis examining the patterns of attention measured by fixations, the patterns of attention measured by acknowledgements also show little difference between the two SA groups. Both bottom and top SA groups show the same approximate patterns of attention.

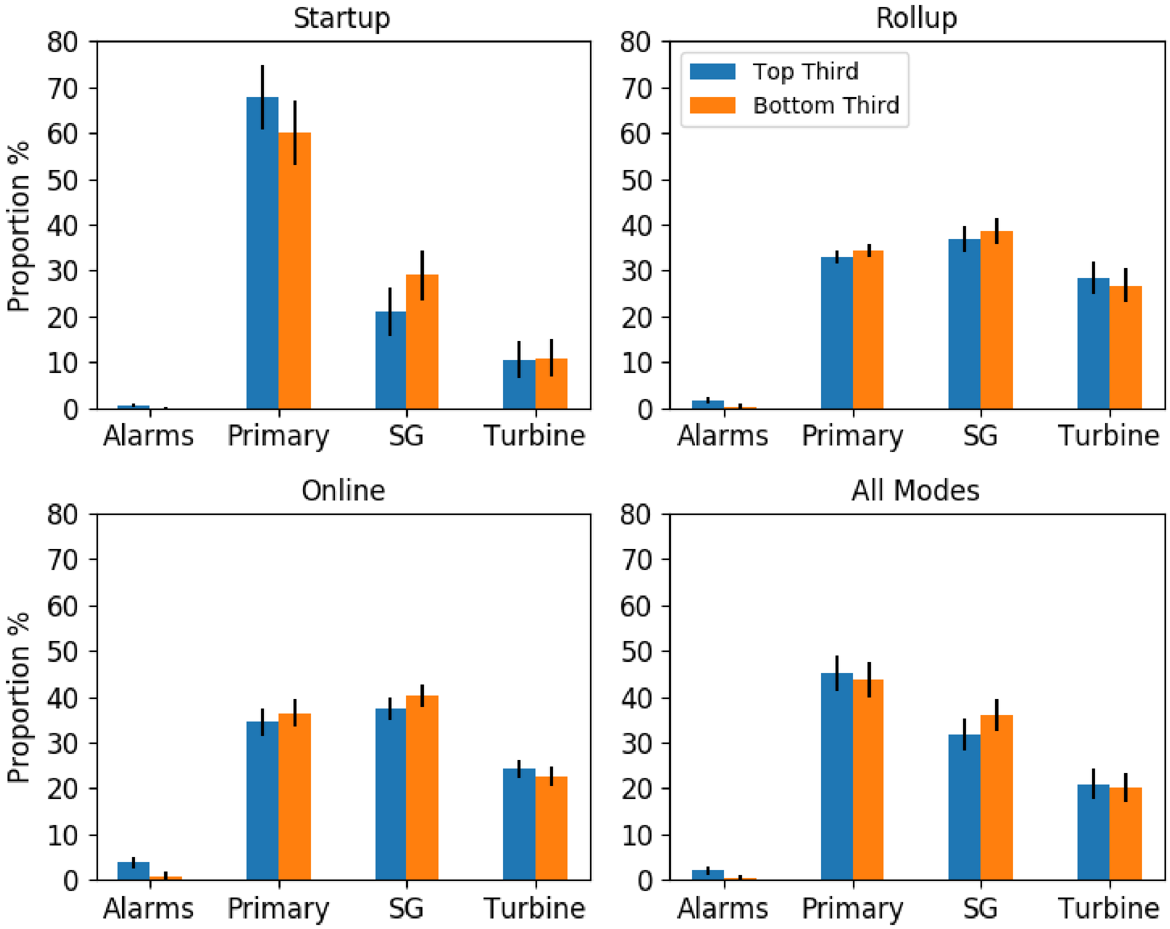


Figure 6.27 Comparison of the top and bottom SA student groups in terms of the patterns of attention measured by the acknowledgement measure within each of the modes of operation.

The patterns of attention measured by acknowledgements were analyzed in separate 2 (performance group) by 3 (mode) repeated measures ANOVA for each of the four interface regions. The results of this analysis can be seen in Table 6.7. Overall participants demonstrated poor SA across the trials, and therefore, the measure simply did not work effectively to correctly capture SA. As a result, the lack of a relationship between attention and SA, which logically should be present, can be accounted for due to the poor ability of the SA measure to effectively assess SA during the trials.

Table 6.7 Results from the ANOVA analyses of the proportions of attention measured by acknowledgements for the top- and bottom-third of participants based on SA scores. No main effect for group was observed indicating both top and bottom performers exhibited similar patterns of attention. The * indicates p < 0.05.

	Source	df	df_{error}	*F*	*p*	η^2	β
Alarms	Mode	2	28	5.51	0.01*	0.28	0.81
	Group	1	14	3.36	0.09	0.19	0.40
	Mode * Group	2	28	2.59	0.09	0.16	0.47
Primary	Mode	2	28	29.55	0.00*	0.68	1.00
	Group	1	14	0.14	0.71	0.01	0.06
	Mode * Group	2	28	0.73	0.49	0.05	0.16
Steam Generators	Mode	2	28	12.64	0.00*	0.47	0.99
	Group	1	14	1.04	0.32	0.07	0.16
	Mode * Group	2	28	0.61	0.55	0.04	0.14
Turbine	Mode	2	28	25.70	0.00*	0.65	1.00
	Group	1	14	0.06	0.81	0.00	0.06
	Mode * Group	2	28	0.10	0.91	0.01	0.06

Mental Model. At the end of the experimental session, the debrief form instructed participants to sketch the system as they remembered it to assess the mental model of the system participants used as they interacted with the simulation. An example of one of the system sketches created by a student participant is shown in Figure 6.28.

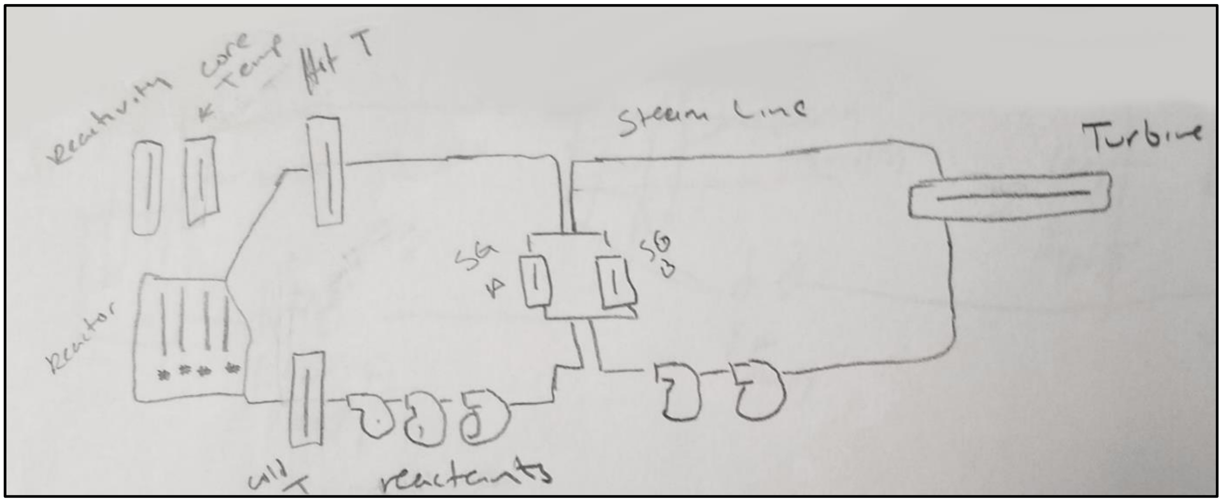


Figure 6.28 System sketch sample depicting the representation of the microworld simulation system reported by a student participant.

The sample sketch of the system drawing shows the common components and indicators the participants typically included in their representations of the system, such as the reactor core, steam generators, and turbine. These sketches were then scored by two different raters to provide a completeness score for each participant, which served as a measure of how comprehensively they drew system components, and an indication with higher scores reflecting more complete system representations and lower scores representing more incomplete system representations.

The two raters independently scored the drawings based on a content analysis approach in which they examined the drawings and judged whether each component and indicator was present or absent (Graneheim & Lundman, 2004; Hsieh & Shannon, 2005). The raters scored the 16 components and 15 indicators as present or absent. The inter-rater reliability was assessed by correlating the present or absent score for each item between the two raters across all participants. The correlations between the two raters for each of the items is shown in Table 6.8. Additionally, the percentage of participants that were scored by both raters as including each item was calculated. Only items that demonstrated an inter-rater reliability of 0.60 or greater were included in the final component and indicator scales. Following the classifications of inter-rater reliability provided by McHugh (2012), the correlations of the included items can be categorized as moderate (0.60-0.79), strong (0.80-0.90), and almost perfect (above 0.90). Additionally, items that had low inclusion rates across participants' drawings, defined by the threshold of 30%, were excluded from the scale. Items that failed to meet both criteria were excluded from further analysis.

Table 6.8 Pearson correlations calculated for each item the raters scored in the system sketches as either present or absent. The items that were included in the component and indicator completeness scale are denoted by *. Correlations that could not be calculated due to only a single rater identifying the item as present are denoted by -.

Components

		Percent of Participants
Item	**Correlation**	**Including Item in Sketch**
*Reactor	0.70	95.59
*SGs	0.80	92.65
*Turbine	0.86	73.53
*Recirculating Pump A	0.80	72.06
*Recirculating Pump B	0.87	67.65
*Recirculating Pump C	0.94	66.18
SG A In	-0.04	4.41
SG B In	-0.04	4.41
SG A Out	0.27	8.82
SG B Out	0.27	8.82
*Feed Water Pump A	0.84	54.41
*Feed Water Pump B	0.94	48.53
Speed	0.75	11.76
Load	0.77	14.71
Bypass	0.84	20.59

Indicators

Item	**Correlation**	
*Reactor Temp	0.78	72.06
*Hot Leg Temp	0.76	58.82
*Cold Leg Temp	0.88	50.00
*Reactivity	0.70	63.24
Main Steam A	0.36	7.35
Main Steam B	0.36	7.35
SG A Level	0.70	4.41
SG B Level	1.00	2.94
Feed Water Flow A	-	1.47
Feed Water Flow B	-	1.47
Primary Flow	-	1.47
Turbine Speed	0.78	27.94
Inlet Press	-	1.47
Power	0.00	11.76
Efficiency	0.80	7.35

The item analysis resulted in a total of eight items used for the component completeness scale and four items used for the indicator completeness scale. Overall, the two raters demonstrated high-levels of agreement on the majority of items. The component and indicator completeness scales demonstrated good internal consistency, as evidenced by Cronbach's alpha values of 0.90 and 0.93, respectively.

The mental model scores were examined within the context of the performance and SA measures. Participants with better mental model scores presumbably have a better understanding of the system, and therefore, they were expected to demonstrate better performance and higher SA. In regard to SA, the mental model scores represent basic elements within the interface and fall under Level I of Endsley's three-level model (1995b). Level I refers to an awareness of individual elements, and therefore, the inclusion of each component or indicators with the sketch provides a metric for participants' awareness of that element. It does not reflect the state or value of that element, so this mental model representation is still distinct from the concept of Level I SA. For the analysis, the component and indicator mental model scores were correlated with revenue, the SAGAT-like SA component value estimate error scores, and the SACRI-like SA trend identification accuracy scores. No significant correlation was observed between either SA measure, or either mental model score. There were significant correlations found between performance, the component completeness score, and the indicator completeness score, as can be seen below in Table 6.9. This is a substantial finding since the SA measures did not correlate with performance or the mental model measure. To the best of the author's knowledge and based on a literature search, no other sketch-based approach has been attempted in a process control simulation or SA domain, and therefore, this represents the first use of this approach. The SA measure appears to have suffered from a floor effect in the student population in which participants demonstrated low-levels of accuracy for the SACRI-like measure and high-error rates for the SAGAT-like measure. In contrast, the mental model measure was able to explain some of the variability in performance, as evidenced by the significant positive correlation. Therefore, the mental model measure shows promise as a method to account for human performance. Specifically, participants that included more components within the system sketch earned more revenue based on the positive correlation that was found. The indicator completeness measure was negatively correlated with performance, which suggests participants that included indicators earned less revenue. The component and indicator completeness measures did not significantly correlate with each other, *r*(25) = 0.00, *p* = 0.50. Initially, this finding was puzzling, but there is one reasonable explanation to this seemingly contradictory finding. The indicators scale contains three temperature indicators for the reactor core, hot leg, and cold leg sections of the primary loop. Since all three temperature indicators are tightly coupled to the thermohydraulics of the reactor core via the position of the control rods, the hot leg and cold legs become redundant to the reactor core temperature indicator. No participants recreated the entire system in their sketch, and therefore, participants included components and indicators that came easily to mind in line with the availability heuristic. The items used attended to and processed most by participants as they interact with the simulation would receive the greatest activation and be recalled the easiest, based on the availability heuristic. Therefore, participants that included these redundant indicators in the system sketch may have been attending to these superfluous indicators. As a result, they did not focus on the more critical elements of the system, and their performance suffered. These results were found within the context of a student population. The next section on the operators will examine whether the mental model measure is effective in explaining an expert population's performance.

Table 6.9 Correlations between the two mental model score measures and performance, SAGAT-like SA component value estimate error measure, and SACRI-like SA trend identification accuracy measure.

Mental Model Score	Measure	*r*	*p*
Component			
	Revenue	0.36	0.04*
	SA Error	-0.32	0.11
	SA Trend	-0.03	0.88
Indicator			
	Revenue	-0.34	0.04*
	SA Error	0.22	0.31
	SA Trend	0.05	0.83

Follow-up Study

The initial microworld study demonstrated a novice student population could in fact learn and control the microworld simulation. Both the eye tracking and acknowledgement measures were sufficiently sensitive to capture the difference in attention patterns across the different modes of operation; however, the eye tracking measure demonstrated greater sensitivity during shorter duration activities. Indeed, the shortest two modes—ready to roll and ready to sync—could not be analyze due to a dearth of acknowledgement data. The other shorter-duration turbine recovery period did demonstrate differences between the acknowledgement and fixation measures. To further examine and validate the use of the acknowledgement measure, a follow-up study was performed to assess the amount of interference generated by the addition of the marker-acknowledgement task. Secondary tasks, such as the freeze probe measure of SA, suffer criticism for intruding upon the primary task (Salmon et al., 2009). Despite this interference, the freeze probe technique is widely employed and provides valuable diagnostic SA information, which suggests that the intrusion is offset by the benefits of the measure. Furthermore, Endsley (1995b) demonstrates that the task-intrusion from the freeze probe is minimal. Since the marker-acknowledgement measure is intended as a means of assessing SA, it is important to determine the amount of interference the measure generates when included with the traditional freeze probe measure. To assess this additional interference, a subset of the original student participants from the initial microworld evaluation were brought back for a second experimental data-collection session. Since this follow-up study examined interference, it was desirable to get high-performing participants so that the issue of simply controlling the simulation was not confounding the amount of interference generated by the additional marker-acknowledgement task. Therefore, a performance threshold of $20k was used to screen out the lowest-performing participants, which resulted in excluding the bottom 20% of participants from the initial study.

Follow-up participants completed the basic process control simulation without the addition of any secondary tasks. A second condition included the traditional freeze probe measure to capture a typical SA study and the interference from the traditional freeze probe technique. A third condition, termed the full condition, included the freeze probe technique with the addition of the marker-acknowledgement task to determine the additional interference due to the marker-acknowledgement task. The eye tracking measure was also used during this follow-up study. Since it is a passive measure of attention-allocation, it does not provide significant interference during the actual trial. It is possible that delays due to technical difficulties could impact the participant, but these are minimal. This further allowed for a comparison in the pattern of attention measured without the marker-acknowledgement task to determine if the pattern of attention differed in the absence of performing this secondary task. Overall, the magnitude of the interference generated by the secondary acknowledgement task was predicted to be minimal. Participants were expected to subjectively rate the difficulty of the conditions in the following order from greatest to least difficulty: full, freeze probe, and basic. The objective performance measures were also predicted to demonstrate this trend with the lowest performance exhibited in the full condition and the highest performance exhibited in the basic condition. No significant difference in the patterns of attention between the acknowledgement-included condition and the other two conditions without acknowledgements was predicted.

Since these follow-up participants also participated in the initial study, the follow-up study served to demonstrate how increased exposure impacts performance on the microworld and allows for further validation of the marker-acknowledgement and SA assessment techniques. Follow-up participants were predicted to demonstrate better performance on the full condition. The full condition in the follow-up study was identical to the trials completed by the participants in the original microworld study, and therefore, serves as a basis for making comparisons due to additional exposure to the microworld. Participants were predicted to demonstrate better objective performance and increased accuracy on the freeze probe responses. Participants in the follow-up study were expected to demonstrate a pattern of attention that is closer to optimal as a result of their increased experience.

Method

Participants. A total of 12 undergraduate psychology students (3 female, 9 male) ranging in age from 18 to 22 years (M = 19.83 years, SD = 1.27 years) participated in the follow-up microworld study. The undergraduate psychology students were comprised of students recruited from the group of participants that participated in the initial study. To qualify for participation in the follow-up study, only participants that exceeded a score of $20,000 in the original study were invited to participate. This criterion eliminated the bottom 20% of participants who participated in the initial study. Participants in the follow-up study were compensated with a monetary payment of $40.

Protocol. The general protocol in the follow-up study was the same as the two previous studies in which participants completed a calibration procedure with the Tobii eye tracking unit prior to each trial and a NASA-TLX paper-based workload measure after completing each trial. The protocol differed from the previous two studies in two important ways. First, the follow-up study did not include any practice trials, since participants had already demonstrated satisfactory performance in the initial study. Second, the follow-up study included different conditions of the acknowledgement and freeze probe secondary tasks. Three conditions were presented to participants following a Latin-square design consisting of the full, freeze probe, and basic conditions, respectively, each being completed twice to provide both turbine- and reactor-fault types. The full condition was identical to the trials student participants performed in the initial study in including the acknowledgement-marker secondary task and three sets of freeze probe questions. The freeze probe condition included only the three freeze probe questions. The basic condition entailed controlling the plant in the absence of either the freeze probe questions or the acknowledgement-marker secondary task.

Results and Discussion

Initial and Follow-up Subjective Performance. Since participants had more experience and practice with the microworld during the second follow-up experimental session, it was expected that they would rate their subject workload as lower than what they reported during the initial experimental session. Only the full condition was used for comparison, since this condition replicated the same experimental setup that participants completed during the initial study. To examine subjective workload, a 2 (session) by 6 (dimension) repeated measures ANOVA on the NASA-TLX self-report scores was performed. A significant main effect for session, $F(1,11) = 11.99$, $p = 0.01$, and dimension, $F(5,55) = 14.60$, $p = 0.00$, was found. A significant interaction between session and dimension was also found, $F(5,55) = 5.19$, $p = 0.00$. As can be seen in Figure 6.29, the students consistently rated each dimension lower, except for physical demand, during the follow-up session in support of the hypothesis that students would demonstrate lower subjective workload. The exception of rating the physical demand higher is the driving force behind the significant interaction that was found between session and dimension. One possible explanation is that participants performed more acknowledgements and control actions during the follow-up experiment. An analysis of these two physical actions showed no significant difference in the number of control actions, *t*(11) = 1.19, *p* = 0.26; however, there was a significant difference in the number of acknowledgements, *t*(11) = 3.75, *p* = 0.00, between the initial (*M* = 187.67, *SD* = 81.97) and follow-up (*M* = 239.50, *SD* = 68.12) sessions. Additionally, participants in the follow-up condition showed higher average reactivity, which translates to a more dynamic system and potentially requiring more rapid responses. As a result, it is not surprising that participants rated physical demand as higher in the follow-up condition, since they performed more acknowledgements and were also required to respond more rapidly due to the control strategy of the reactor they used. The participants also rated performance lower, which is counter to what was predicted. With more experience, participants were predicted to rate their performance better. The lower performance ratings can be attributed to a better understanding of the maximum score that can be achieved for revenue. Participants in the follow-up study became aware of the possible high-score they could have achieved, and therefore, rated their own performance lower within that context. Indeed, nearly all the participants asked for the current high-score, since they were highly motivated to beat it.

The knowledge of this high-score appears to be reflected in the subjective ratings of performance.

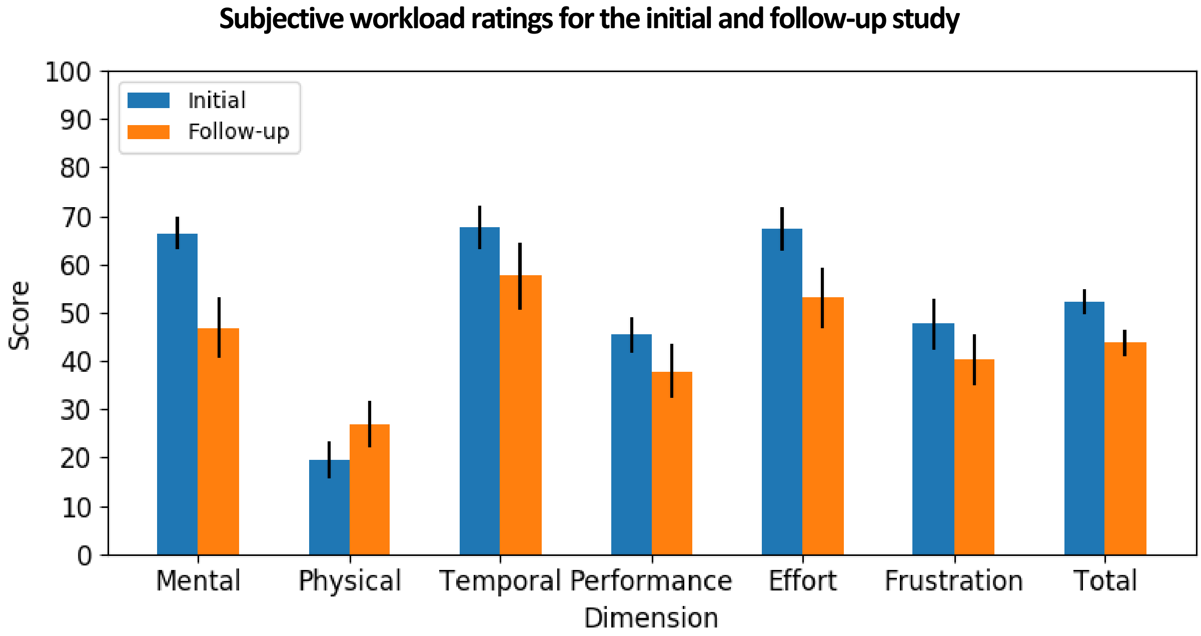


Figure 6.29 A comparison of the subjective ratings of workload reported for each dimension of the NASA-TLX between the initial and follow-up experimental studies. Error bars indicate standard error for each mean.

Secondary Task Subjective Performance. To assess the impact of the secondary tasks in the freeze probe and full condition on the subjective workload, the NASA-TLX scores for each condition were compared. A 3 (condition) by 6 (dimension) repeated measures ANOVA on the NASA-TLX self-report scores was performed. A significant main effect for dimension was found, *F*(5,55) = 6.38, *p* = 0.00. No significant main effect for condition was found, *F*(2,22) = 1.96, *p* = 0.17. No significant interaction between condition and dimension was found, *F*(10,110) = 1.55, *p* = 0.13. The main effect for dimension is not overly interesting since this is purely due to the design of the NASA-TLX as a measure of different workload dimensions. The lack of a main effect for condition provides some evidence of the perceived equivalency between the three conditions. The effect size for condition was quite small, η^2 = 0.01, and the observed power was 0.86, which together indicates the lack of a statistical difference is likely not due to a small sample size, but potentially due to a true lack of difference between the conditions. Furthermore, from a qualitative perspective, the workload measure is similar for the six dimensions across the different conditions, as shown in Figure 6.30. It is likely that the conditions do not differ in a meaningful way based on the results from this data, which provides support that the addition of the freeze probe and the freeze probe with the acknowledgement task are not perceived as imposing a greater workload by the participants.

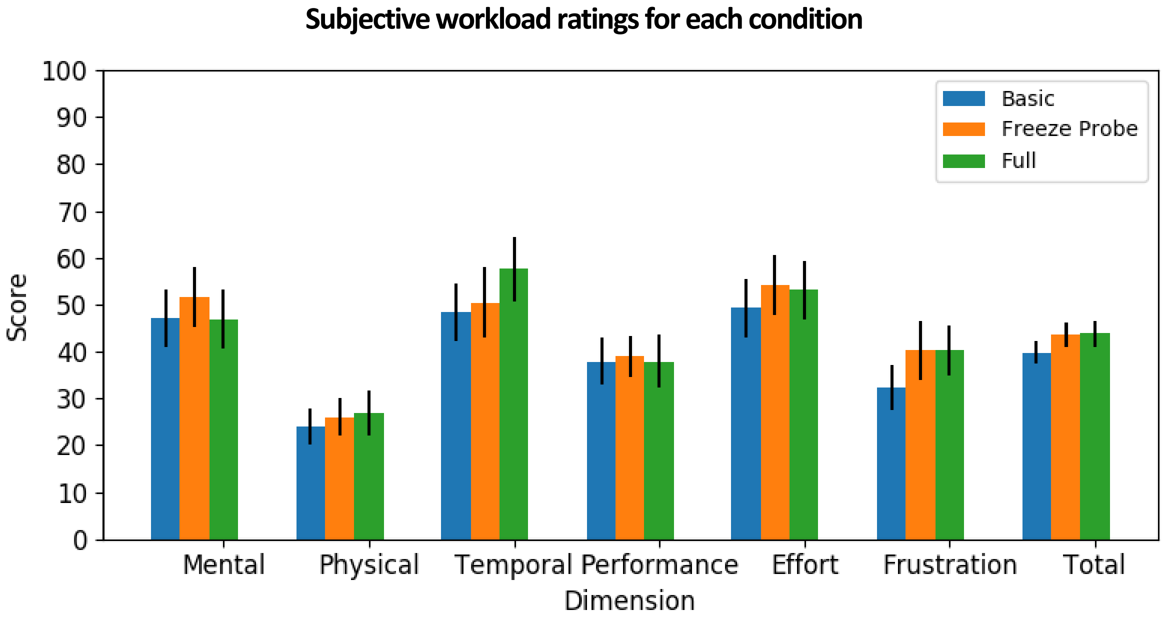


Figure 6.30 A comparison of the subjective ratings of workload reported for each dimension of the NASA-TLX between the three experimental conditions in the follow-up study. Error bars indicate standard error for each mean.

Initial and Follow-up Performance Comparisons. In this second experimental session, participants completed two full condition trials identical in nature to the trials within the first experimental session. Only these full condition trials were included in comparisons made between the initial and follow-up sessions. To examine the impact of additional exposure on performance, performance in the follow-up study was compared to performance in the initial study. Only students that participated in both were included in the analysis. The performance measures from the initial study were collapsed across the four trials to yield representative performance measures for the initial study. Since this was the second experimental session and participants had experience completing the trials during the initial study, participants were predicted to demonstrate better performance in the follow-up session than the initial session.

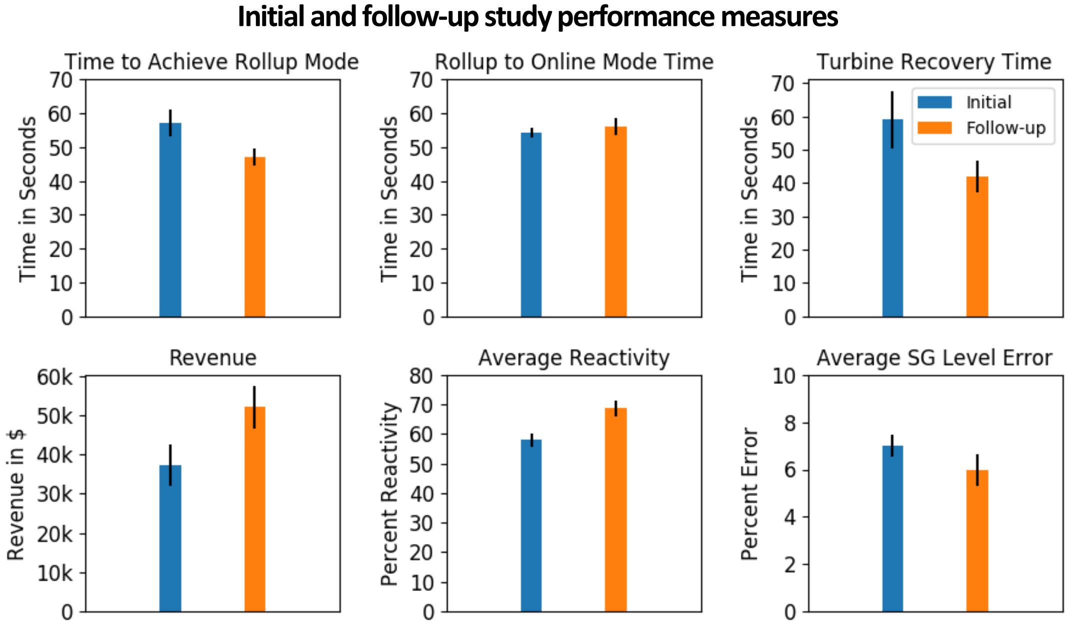


Figure 6.31 Performance measures comparisons between the initial and follow-up study for the timing, revenue, and component control error measures of performance. Error bars indicate standard error for each mean.

A graphical depiction of the results of the performance measure analyses can be seen in Figure 6.31. There was a significant difference in the revenue earned between the first (*M*

= 37.2k, *SD* = 18.0k) and second (*M* = 49.9k, *SD* = 20.7k) experimental sessions, *t*(11) = 7.33, *p* = 0.02. A significant difference in the time to achieve the turbine rollup mode of operation was found between the first (*M* = 57.25, *SD* = 14.58) and second (*M* = 42.91, *SD* = 13.68) experimental sessions, *t*(11) = 8.03, *p* = 0.02. However, there was no significant difference in elapsed time between the turbine rollup and online mode of operation between the first and second experiments, *t*(11) = 1.31, *p* = 0.28. Reactivity and steam generator level average error—component-specific measures of performance—were also examined between the first and second sessions. A paired-sample t-test showed a significant difference between percent reactivity in the first (*M* = 58.02, *SD* = 12.18) and second (*M* = 70.56, *SD* = 13.19) experimental sessions, *t*(11) = 22.63, *p* = 0.00. This difference in reactivity mirrors the difference in revenue generated since with additional heat input to the system through more reactivity, the students could produce more steam—and ultimately more power and revenue—over the trial. The difference in reactivity reflects an improvement in control strategy in which the participants were performing more optimally with the reactor closer to the optimal 100% power level. No significant difference was observed between average steam generator level deviations in the first (*M* = 7.70%, *SD* = 2.34%) or second (*M* = 6.00%, *SD* = 3.54%) experimental sessions, *t*(11) = 1.90, *p* = 0.09. There was no effect of trial observed on steam generator level deviations during the initial experimental session, and participants were already maintaining the deviation within 10% of the optimal 50% fill level during the first session. Taken together, the lack of improvement during either the initial or follow-up study indicates performance had already plateaued to levels in which the measure was not sufficiently sensitive to detect any significant performance differences. Thus, the strategy for controlling the steam generators likely did not change in a meaningful way between experimental sessions.

Secondary Acknowledgement Task Performance Comparisons. The full, freeze probe, and basic conditions were compared to determine the impacts of the secondary tasks on performance as measured by revenue, the times to achieve modes and recover from the turbine fault, and controlling component errors. Revenue generated for each condition was examined across conditions with a repeated measures ANOVA. No significant main effect for condition on revenue earned was observed, *F*(2,22) = 1.90, *p* = 0.17, which provides some supporting evidence that the secondary task did not interfere significantly with the primary simulation task.

In addition to total revenue earned, the timing measures of performance including the time to achieve turbine rollup, the elapsed time between turbine rollup and the online mode of operation, and recovery times from the scripted turbine fault were examined across the three conditions with three separate repeated measures ANOVAs. A significant main effect for condition on the time to achieve turbine rollup, *F*(2,22) = 3.88, *p* = 0.04, was found. There was no significant effect for condition on the time to achieve the online mode of operation, *F*(2,22) = 0.83, *p* = 0.45. No main effect for condition on recovery time, *F*(2,22) = 1.83, *p* = 0.18, was observed either. The component-specific measures of performance for reactivity and steam generator level deviations across the three conditions were analyzed with two separate repeated measures ANOVAs. A significant main effect was observed for condition on reactivity, *F*(2,22) = 3.67, *p* = 0.04, as shown in Figure 6.32. A Fischer's LSD test showed the basic condition exhibited significantly higher reactivity than the full condition, while the freeze probe condition was not significantly different from either of the other two conditions. No significant main effect was observed for condition on steam generator level deviations, *F*(2,22) = 1.41, *p* = 0.27.

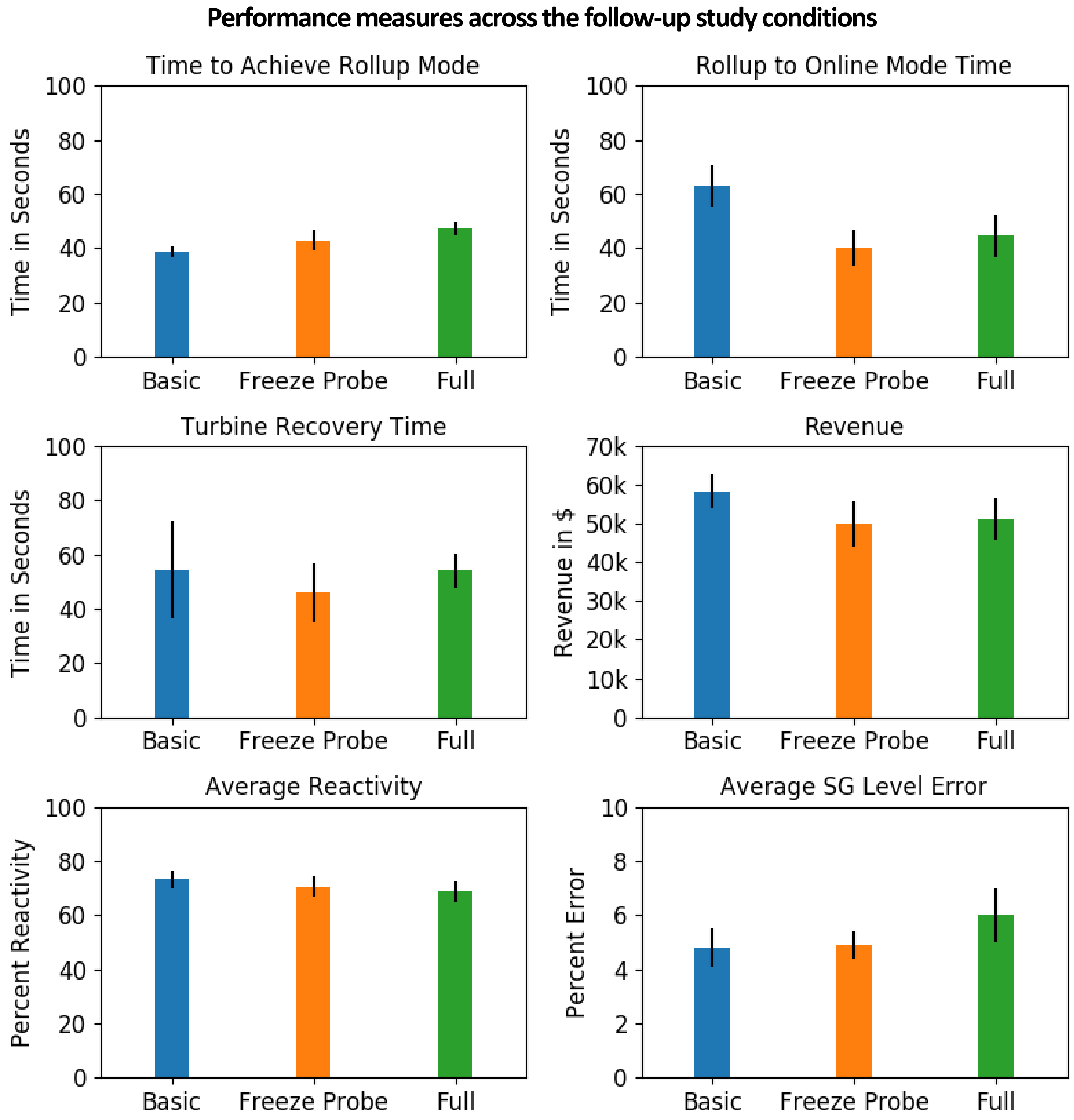


Figure 6.32 Comparisons of timing and maintaining specific-component values performance measures between the basic, freeze probe, and full conditions in the follow-up study. Error bars indicate standard error for each mean.

In line with the analysis of the students performed during the initial study, the reactivity measure of performance demonstrates greater sensitivity and yields a significant effect for condition in the follow-up study. Since a secondary task should always impose some performance decrement, it is probable that performance does in fact differ to some degree between conditions, but the revenue and timing measures may simply be too blunt to detect these differences. However, since this interference can only be detected by the more sensitive measures, the amount of interference is relatively small and does not impact performance sufficiently to preclude its use. Ultimately, researchers must determine whether this potential additional secondary task intrusion is offset by the benefits of capturing attention patterns with the measure within the context of the goals of the research. For this particular line of research, which has an emphasis on attention, the benefits outweigh the drawbacks of the secondary task interference.

Initial and Follow-up Situation Awareness Comparisons. Students in the follow-up study were expected to demonstrate better SA due to the additional practice and experience with the microworld simulation. The SA measures in the initial study demonstrated overall poor levels of SA, and as a result, the measures' sensitivity was lacking. It is possible that with more experience, participants will be able to achieve better levels of SA, similar to what an experienced operator would exhibit. To examine any improvement, SA measured by the SAGAT-like measure, component value estimate error, and the SACRI-like measure, trend identification accuracy in the initial study was compared with the full condition in the follow-up study. The freeze probe only condition from the follow-up study was not included in this analysis because it did not contain the same experimental setup, and therefore, would not be appropriate for comparison. The SA probe responses were examined by region and period. The examination across regions supports comparisons between groupings of functionally and physically collocated components to determine how SA fluctuates across regions of the interface. The examination across periods affords comparisons between SA of the entire system, or SA collapsed across all components in all regions, at different time-points during the simulation to understand how SA fluctuated over the course of the trial. A 2 (session) by 3 (period) by 3 (region) repeated measures ANOVA was performed on component value estimate response error. No significant main effect for session, *F*(1,11) = 3.80, *p* = 0.08, or period, *F*(2,22) = 1.38, *p* = 0.27, was found. A significant effect for region was found, *F*(2,22) = 14.88, *p* = 0.00. No significant interactions were found.

SAGAT-like component value estimates error for initial and follow-up studies across regions

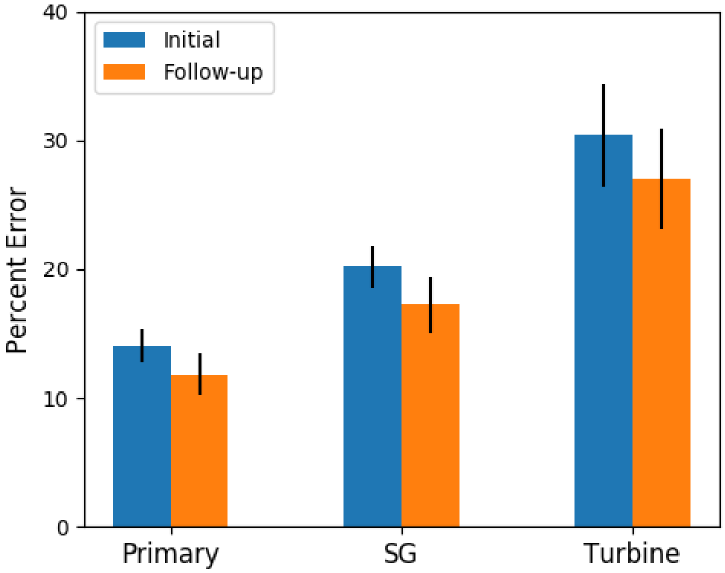


Figure 6.33 A comparison of SA component value estimate response error between the initial and follow-up studies to show the main effects for session and region. Error bars indicate standard error for each mean.

The component value estimate errors did not demonstrate elevated SA in the follow-up study, as can be seen in Figure 6.33. Due to the lack of sensitivity for the component value error SA measure in the initial study, this was not entirely a surprise. The measure may simply not be sensitive enough to detect the improved SA participants in the follow-up study relied on to achieve better performance. It is also possible that participants did not have sufficient time to further develop their mental representation of the system for the SA measure to capture a significant difference. Another explanation is that students achieved their best SA and maintained that during the second session. Their SA may have plateaued and they did not meaningfully improve upon their mental representation of the system from what they had achieved during the initial study. This does not mean that higher SA could not be achieved through other means, such as additional training and instructional material that could improve the students' mental representation of the system. The main effect for region provides further evidence that the turbine region of the interface is the most complicated region of the interface as predicted and as also demonstrated by SA scores from the initial study. The significant effect for region provides additional support for the use of a SAGAT-like, specific component value estimate probe measure of SA (Endsley, 1995b) in a microworld setting to provide diagnostic information concerning how SA differs between groupings of system components. The ability to detect challenging aspects of the system can be informative for design, which lends itself to eventually examining new interface designs and quantifying how they impact SA (Hogg et al., 1995).

In addition to the component value estimate response error measure of SA, trend identification accuracy between the initial and follow-up studies was also compared. A 2 (session) by 3 (region) by 3 (period) repeated measures ANOVA was performed on correct trend identification accuracy. A significant main effect for region, *F*(2,22) = 6.74, *p* = 0.01, and period, *F*(2,22) = 16.55, *p* = 0.00, was found. No significant main effect for session, *F*(1,11) = 4.69, *p* = 0.05, was found, but a significant interaction between region and period was found, *F*(4,44) = 6.71, *p* = 0.00. Similarly to the component value estimate error SA measure, no significant increase in correct trend identifications was found for the follow-up condition as can be seen in Figure 6.34. The measure of SA did not work well. Due to its lack of sensitivity, it was not able to capture any differences in SA during the follow-up session. The lack of improvement could also be attributed to the same rationale mentioned for the component value estimate error section above in which participants had already achieved their best SA during the initial session or the additional exposure simply did not translate to meaningful improvements for SA.

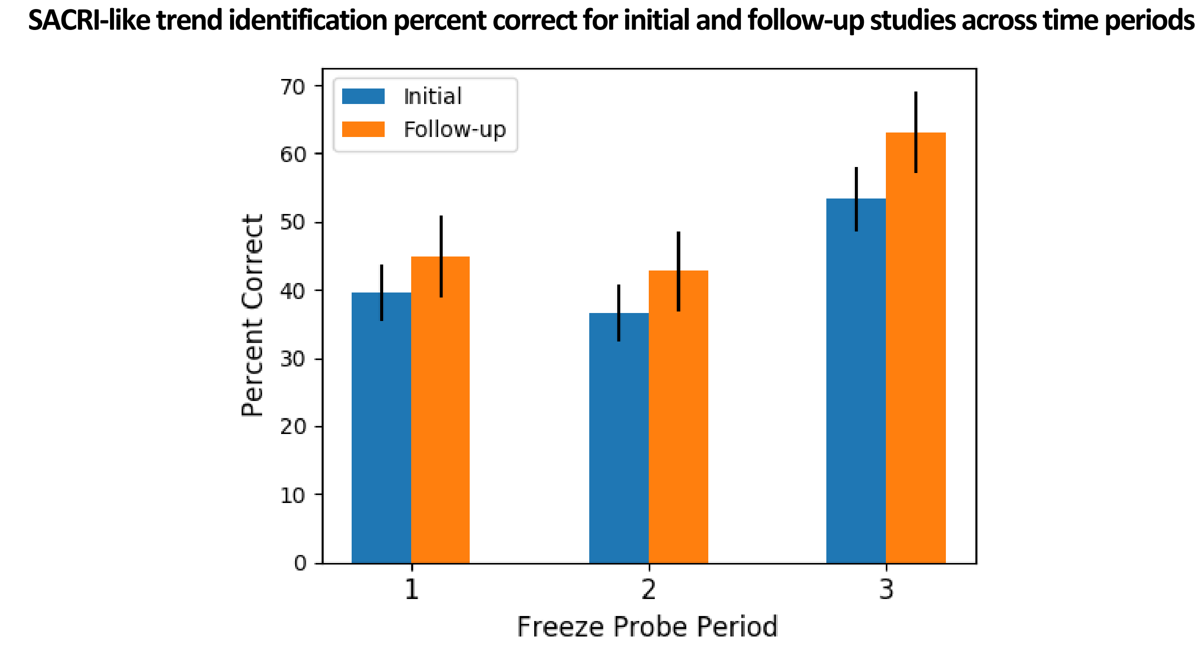

SACRI-like trend identification percent correct for initial and follow-up studies across time periods

Figure 6.34 A comparison of SA correct trend identification response accuracy between the initial and follow-up studies to show the main effects for session and region. Error bars indicate standard error for each mean.

It is important to note that though the SACRI-like measure did not detect differences between sessions, it was capable of detecting significant effects for region and period, which provides supporting evidence for the future use of the SACRI-like measures of SA in microworld settings. The SACRI has been used in full-scope simulations to differentiate levels of SA between operators using different interfaces (Hogg et al., 1995); here, we have used a similar version in the microworld that was capable of generating significant effects based on regions of the interface and time-periods within the trials. Establishing the measure can be used with even limited success in a different setting, such as a microworld, is an interesting result.

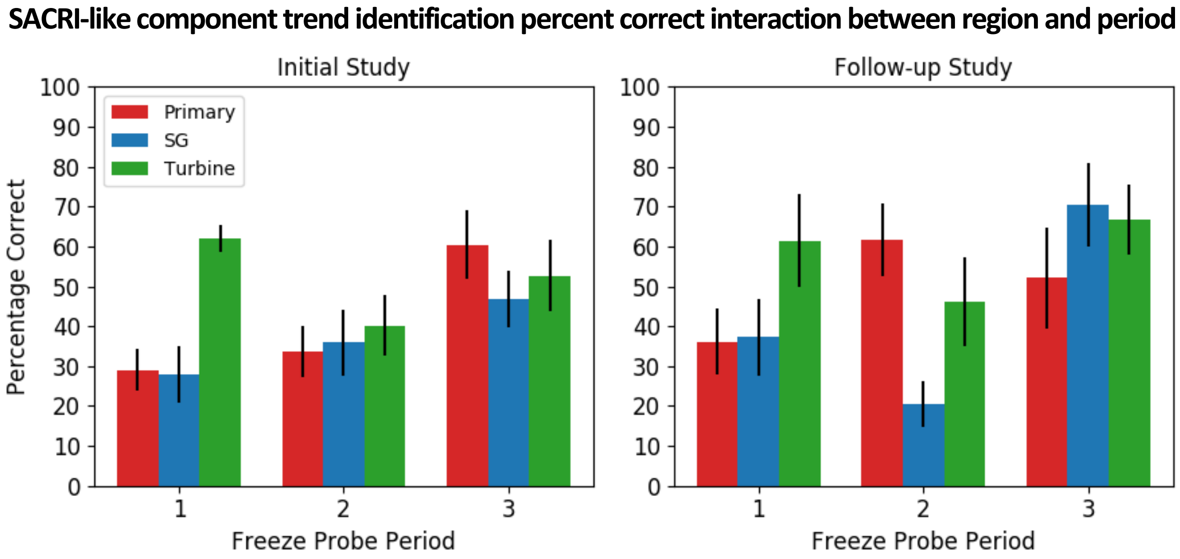


Figure 6.35 A comparison of SA correct trend identification response accuracy between the initial and follow-up studies to show the interactions between session, period, and region. Error bars indicate standard error for each mean.

The three-way interaction between session, period, and region is shown in Figure 6.35. The turbine region demonstrated a pattern for correct trend identification in which participants exhibited the highest accuracy during the first period, the lowest accuracy in the second period, and high-accuracy again in the third period. The reason for this pattern is that during the first period, the turbine is not in operation, and therefore, it becomes simple to report an unchanged state to turbine region freeze probes. Knowledge of the unchanged state reflects achieving Level II SA in which participants are integrating the various components into a mental model. As a result, participants are capable of identifying that during the startup mode of operation, the turbine is offline and its related components will remain largely unchanged. The second period represents the turbine undergoing the greatest fluctuations, since the turbine has recently been synced to the grid. As a result, the steam inlet pressure and power-produced trends are more challenging for participants to correctly identify. The third period also demonstrated high-levels of correct trend identification within the turbine region. This high-accuracy can be attributed to the steady-state in which the turbine and the related parameters are operating within while in the online power-producing mode of operation. As a result, the components do not fluctuate as much as just after syncing to the grid, and therefore, it is easier to identify trends. In regard to session and the turbine region interactions, the accuracy in the final period was higher for the follow-up study than in the initial study as would be expected since students are more familiar with the process and presumably have a better understanding of the underlying simulation model. This additional experience translated to increased accuracy in trend identification in the follow-up study. This provides some evidence that the SA measures themselves were not invalid, but rather participants must have a certain amount of experience for the measure to operate. Initially, participants may have found the probe interrupts too challenging, while after more experience, they were better able to respond, thereby causing the measure to become more diagnostic.

The steam generator region of the interface follows a somewhat similar pattern as the turbine region. During the initial study, participants demonstrated the poorest trend identification accuracy during the first period, while in the second study, participants demonstrated the poorest trend identification during the second period. This may reflect a shift in the understanding of the system, which has been gained during the follow-up study. Participants have gained the understanding that during the startup mode of operation, the steam generator levels will decrease as they lose water to the steam production process unless they are manipulated. Furthermore, any manipulations are now represented more accurately in the participants' mental model, and therefore, during the first period, participants demonstrate better trend identification accuracy. The final period is administered during an online mode of operation in which participants have achieved a relatively steady-state within the steam generators, and therefore, they are able to more accurately identify the trends. In the last period, participants demonstrated the highest trend identification accuracy within the steam generator region in both the initial and follow-up studies, but the follow-up study is significantly more accurate, which reflects an improvement in understanding the system and better SA.

The primary region trend identification accuracy did not differ between the initial and follow-up studies in period one or two. Trend identification accuracy for the primary region in period two was significantly greater in the follow-up study than in the initial, which suggests participants became better at understanding the fluctuations within the primary region during the follow-up study.

Secondary Acknowledgement Task Attention Comparisons. Since the acknowledgement measure is a secondary task, which potentially interferes with the primary task, patterns of attention were compared between conditions to determine if the acknowledgement task altered the participants' attention. Ideally, the acknowledgement task would not alter patterns of attention, since they were designed to function within the natural pattern of attention required for the primary process control task. It was predicted that the additional acknowledgement task in the full condition would not alter the pattern of attention. To assess any differences in the proportions of attention in each of the four interface regions between the basic, freeze probe, and full conditions, separate repeated measures ANOVAs were performed. To increase power, an alpha of 0.10 was used in place of the standard 0.05 to determine significance.

Table 6.10 Statistical results for the analyses of the proportions of attention, assessed with the eye tracking fixation measure, within each region between the three experimental conditions in the follow-up study. The * indicates p < 0.10.

	df	**df$_{error}$**	***F***	***p***	**η^2**	***Power***
Alarms	2	22	2.89	0.08*	0.21	0.64
Primary	2	22	0.58	0.57	0.05	0.22
Steam Generators	2	22	0.55	0.59	0.05	0.22
Turbine	2	22	5.61	0.01*	0.34	0.89

There were no significant main effects for condition on the proportion of attention in the primary and steam generator regions, as can be seen in Table 6.10. There was a significant effect for condition on the proportion of attention in the alarms and turbine regions between the three experimental conditions, as can be seen in Figure 6.36. A post hoc Tukey test showed the proportion of attention to the alarm region was significantly higher in the freeze probe condition than either the basic or full conditions, which were not significantly different from each other. There is no clear explanation for this pattern of results since the full and basic conditions are not significantly different from each other. A post hoc analysis for the turbine region showed the basic condition significantly differed from the full condition for the proportion of attention. In the absence of the acknowledgement task, participants forgo the additional secondary task, which frees them to scan the interface more, and as a result, it appears they checked the turbine more in the basic condition than they did in the full condition with the acknowledgement task.

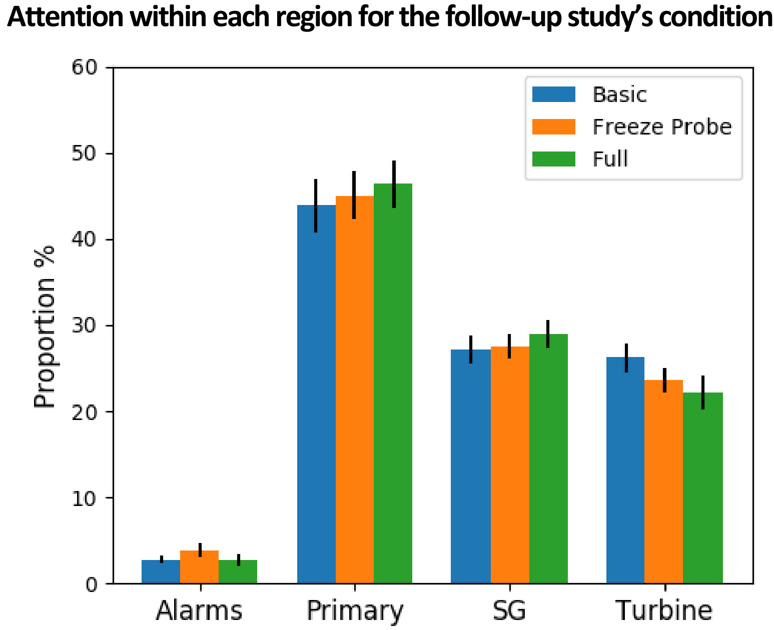


Figure 6.36 A comparison of the proportions of attention, assessed with the eye tracking fixation measure, within each region for the basic, freeze probe, and full conditions in the follow-up study. Error bars indicate standard error for each mean.

It is important to note that the lack of significant main effects observed for condition does not affirm that there is no difference in patterns of attention. Indeed, there is a difference as indicated by the increased attention directed to the turbine region in the basic condition as opposed to the full condition. However, in general, the overall patterns of attention suggest that there may be little meaningful difference between the conditions. Qualitatively, the patterns of attention assessed in the conditions follow the same general pattern in which alarms receive very little attention, the primary region receives the most, the steam generator region receives less than the primary, and the turbine receives less than the steam generators. Collectively, this evidence suggests it is reasonable to assume that the patterns of attention may not meaningfully differ between conditions. The potential interference from the addition of the acknowledgement task, reflecting as a slight change in the attention pattern, does not meaningfully change the way participants distribute attention across the display; thus, the acknowledgement measure remains a worthwhile substitute when eye tracking data is not feasible due to technical limitations or cost (Boring et al., 2012a; 2012b; 2013; Kovesdi et al., 2015). It is up to the researcher interested in using this measure to determine whether a small alteration to the pattern of attention is offset by the value of assessing attention and inferring SA beyond what is available with the traditional freeze probe methodology.

Initial and Follow-up Study Attention Comparisons. Participants in the follow-up session were expected to demonstrate a more optimal pattern of attention. Following the rationale of the SEEV model, participants with more experience can more efficiently attend to pertinent interface elements as they interact with the simulation, and therefore, their patterns of attention were expected to be more optimal. Only the trials in the full condition were included, since these trials are equivalent to the trials participants encountered during the first experimental session. First, the proportions of attention within each region as measured by fixations were analyzed. A 2 (session) by 3 (mode) repeated measures ANOVA was performed for each of the four interface regions. The results of these analyses can be seen in Table 6.11. No main effects for session were found, indicating that the pattern of attention as measured by the proportions of fixations remained unchanged between the first and second experimental sessions. Patterns of attention within each mode can be seen in the left column of Figure 6.37.

Table 6.11 Results from the ANOVA analyses show the proportion of fixations for participants in the initial and follow-up experimental sessions. No main effect for session was observed, indicating the pattern of attention did not differ between sessions. The * indicates p < 0.05.

	Source	df	df_{error}	*F*	*p*	η^2	β
Alarms	Session	1	22	1.54	0.24	0.12	0.21
	Mode	2	22	4.55	0.02*	0.29	0.71
	Session * Mode	2	22	0.59	0.56	0.05	0.14
Primary	Session	1	22	2.42	0.15	0.18	0.30
	Mode	2	22	45.75	0.00*	0.81	1.00
	Session * Mode	2	22	1.00	0.38	0.08	0.20
Steam Generators	Session	1	22	2.24	0.16	0.17	0.28
	Mode	2	22	16.57	0.00*	0.60	1.00
	Session * Mode	2	22	0.15	0.86	0.01	0.07
Turbine	Session	1	22	0.49	0.50	0.04	0.10
	Mode	2	22	32.15	0.00*	0.75	1.00
	Session * Mode	2	22	1.96	0.16	0.15	0.36

Attention across regions measured by acknowledgements and fixations for students and operators

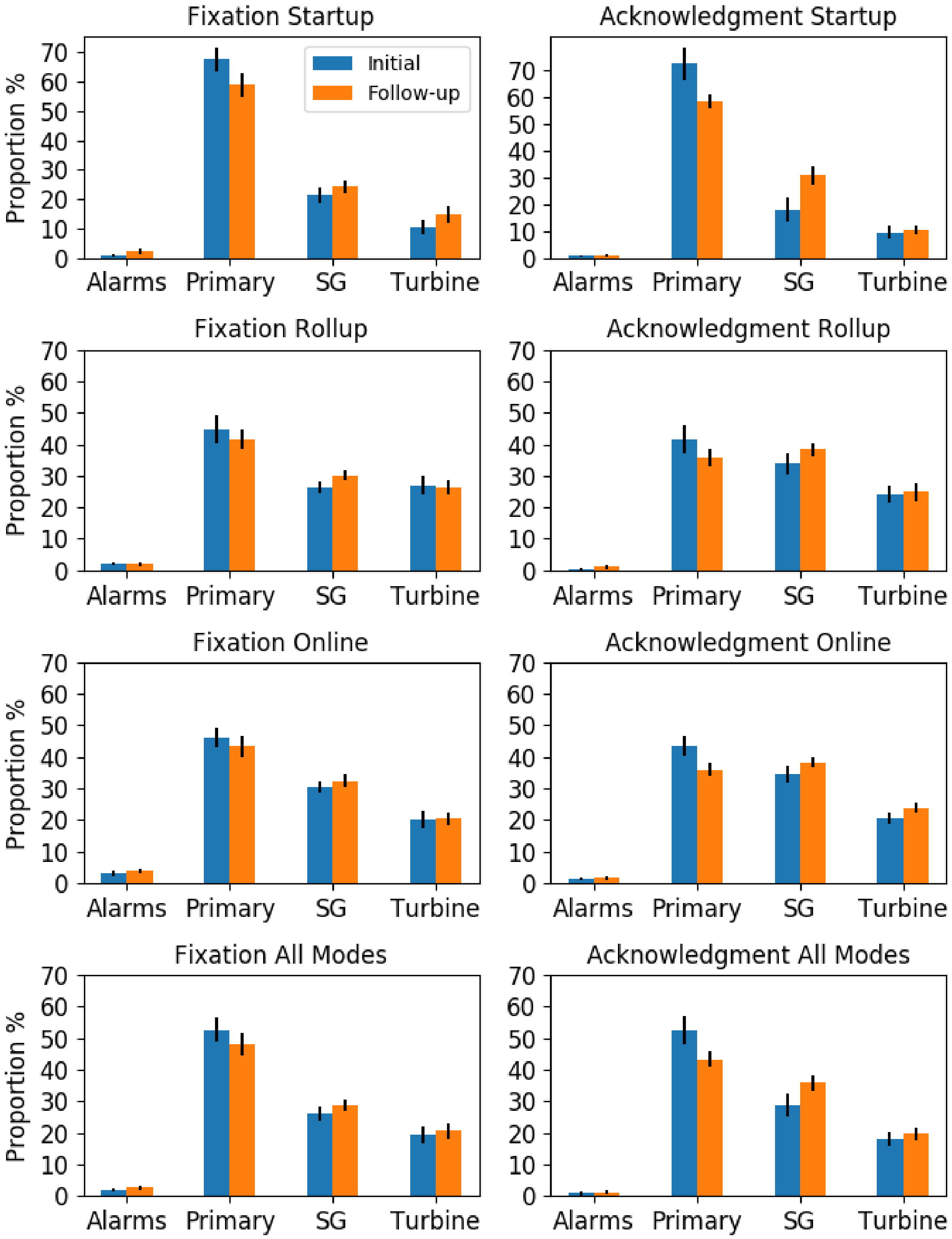


Figure 6.37 A comparison of the proportions of attention within each mode for students in the initial and follow-up studies. Error bars indicate standard error for each mean.

The proportions of attention within each region as measured by acknowledgements were analyzed using a 2 (session) by 3 (mode) repeated measures ANOVA for each of the four interface regions. The results of these analyses can be seen in Table 6.12.

Table 6.12 Results from the ANOVA analyses show the proportion of acknowledgements for participants in the initial and follow-up experimental sessions. A main effect for session was observed, indicating the pattern of attention did not differ between sessions. The * indicates p < 0.05.

	Source	df	df_{error}	*F*	*p*	η^2	β
Alarms	Session	1	22	0.71	0.42	0.06	0.12
	Mode	2	22	1.41	0.27	0.11	0.27
	Session * Mode	2	22	0.17	0.84	0.02	0.07
Primary	Session	1	22	14.38	0.00*	0.57	0.93
	Mode	2	22	33.82	0.00*	0.75	1.00
	Session * Mode	2	22	1.06	0.37	0.09	0.21
Steam Generators	Session	1	22	11.54	0.01*	0.51	0.87
	Mode	2	22	14.17	0.00*	0.56	1.00
	Session * Mode	2	22	3.23	0.06	0.23	0.56
Turbine	Session	1	22	2.20	0.17	0.17	0.27
	Mode	2	22	39.12	0.00*	0.78	1.00
	Session * Mode	2	22	0.71	0.50	0.06	0.16

The analysis for the proportion of attention measured by the acknowledgements did demonstrate main effects for session in the primary and steam generator regions. This is a puzzling result since the fixation pattern of attentions did not show any significant main effects for the primary or steam generator regions. One possible explanation for this discrepancy is that the pattern of attention did not change, but instead participants became more efficient in the secondary task specifically within those regions. These two regions represent where attention is allocated at the highest proportions, as measured by fixations. Therefore, it is possible that the response behavior was influenced by participants directing their acknowledgement efforts within these regions. They still directed attention elsewhere as they performed the task, but they maintained their cursor over these regions to access the controls located in these regions, which may have skewed the proportions of acknowledgement measured within these regions. This represents a shortcoming of acknowledgement measure since it is a secondary task and does require additional effort. If this effort skews the pattern of acknowledgements, then the measure will incorrectly assess patterns of attention.

Limitations. The analyses discussed in the previous section pertain to the basic simulation, simulation with freeze probes, and full simulation with freeze probes and the acknowledgement task. A fourth condition consisting of the marker-acknowledgement task was considered but ultimately ruled out for the following reasons. To balance the number of trials participants were exposed to against the original six (two practice trials and four experimental trials) from the initial study, six trials were used in the follow-up, which only allowed for three conditions to be examined (i.e., each condition contained both a reactor and turbine trip fault). Furthermore, the structure of the conditions in this arrangement did afford for comparisons to be made in terms of any additional interference generated by the marker-acknowledgement task over the traditional freeze probe measure of SA. The marker-acknowledgement measure was never intended as a replacement for the freeze probe, but rather as a means to augment SA assessment and replace eye tracking when feasible. There are few instances in which a researcher would likely forgo administering the freeze probe technique, and therefore, its inclusion with the acknowledgement marker represents an ecologically valid use of the measure, and thus, a suitable condition to examine further.

Future research can provide additional quantification of the interference by also including this fourth condition.

Operator Expertise Study

The primary objective of the operator expertise study centered on comparing an expert population with experience in process control to the novice population sampled in the initial microworld evaluation study. This objective relates primarily to the concept of discount usability (Nielson, 1989). By using inexpensive and more accessible student populations, more researchers are afforded the opportunity to conduct research in the complex process control domain than would otherwise be possible. To determine whether students can serve as a substitute population for operators, the students must first be compared to operators to ensure that student-based experiments using process control microworld studies can generate meaningful conclusions that can be generalized to the expert population. To this end, the operators' performance, SA, and attention distributions were compared to that of the students from the initial study. Characterizing the differences between students and operators is important to identify what aspects can be generalized and what aspects should be avoided. Operators were predicted to demonstrate better performance, better SA, and a different pattern of attention than what students exhibited. Furthermore, examining an expert population with process control experience serves as additional validation for the microworld platform as a research tool to study complex process control.

Method

Participants. A total of eight university-employed steam plant operators (1 female, 7 male) ranging in age from 25 to 46 (M= 33.88 years, SD= 6.94 years) participated in the microworld expertise study. Experience working as a licensed operator ranged from 1.5 to 17 years (*M* = 8.36, *SD* = 6.36). All of the operators reported experience in process control-specific areas to steam production, since this is the central process occurring at the university steam plant, which provides steam for climate control and hot water on campus. Based on the debrief form, the operators reported experience with Siemens and Yocagawa control systems and their associated HMIs. These systems represent industry standards and are also commonly used in the nuclear process control domain within the main control room. As such, the operators had experience working with similar interfaces while controlling an analogous steam production process following the same fundamental principles found within the nuclear process control domain. Furthermore, one of the operators reported experience working at a NPP in the south west, though this operator was not employed as a licensed control room operator and performed maintenance on equipment while working at the plant. It is important to note that these were not in fact nuclear operators, but operators in a related process control plant. These operators received much less training on a more limited system, and therefore, generalizing the findings to nuclear operators is premature. However, these operators do possess experience and expertise in complex process control and faithfully serve as a basis for comparing novice and expert populations using the microworld simulation.

Protocol. The general protocol described previously in the method section of the initial microworld evaluation study was followed for the operator expertise study. After operators gave informed-consent, they experienced the same training as the undergraduate students in study one, and then proceeded to complete trials with the Rancor Microworld. At the start of each trial, operators performed the calibration procedure with the eye tracking system. Upon completing each trial, operators completed the NASA-TLX questionnaire. After completing the six trials, the operator completed the general debrief form and an additional operator-specific debrief form to capture aspects of their experience and expertise gained from working in the process control domain.

Results and Discussion

Subject Performance. Operators were predicted to report lower subjective workload than students since they have previous experience with similar process control at the steam plant. As a result, they should have the requisite knowledge to control the system more easily, and therefore, the simulation should pose less of a challenge and workload should remain lower. To analyze operator and student responses to the NASA-TLX measure of workload, a mixed design ANOVA was performed with a between subjects factor of expertise and a within subjects factor of dimension. Significant main effects for expertise, *F*(1,32) = 306.96, *p* = 0.00, and dimension, *F*(5,160) =27.33, *p* = 0.00, on subjective workload were found. No significant interaction between expertise and dimension was found, *F*(5,160)

= 1.87, *p* = 0.10. As with the previous two studies, the significant dimension effect is not of much interest, since this more reflects the different constructs each dimension measures. The main effect for expertise supports the hypothesis with the operators consistently reporting lower scores than students for each of the six dimensions, as can be seen in Figure 6.38. A Fischer's LSD post hoc analysis showed the means between students and operators were not significantly different in the performance dimension, but differed significantly in all other dimensions with operators reporting significantly lower workload on each of the other five dimensions.

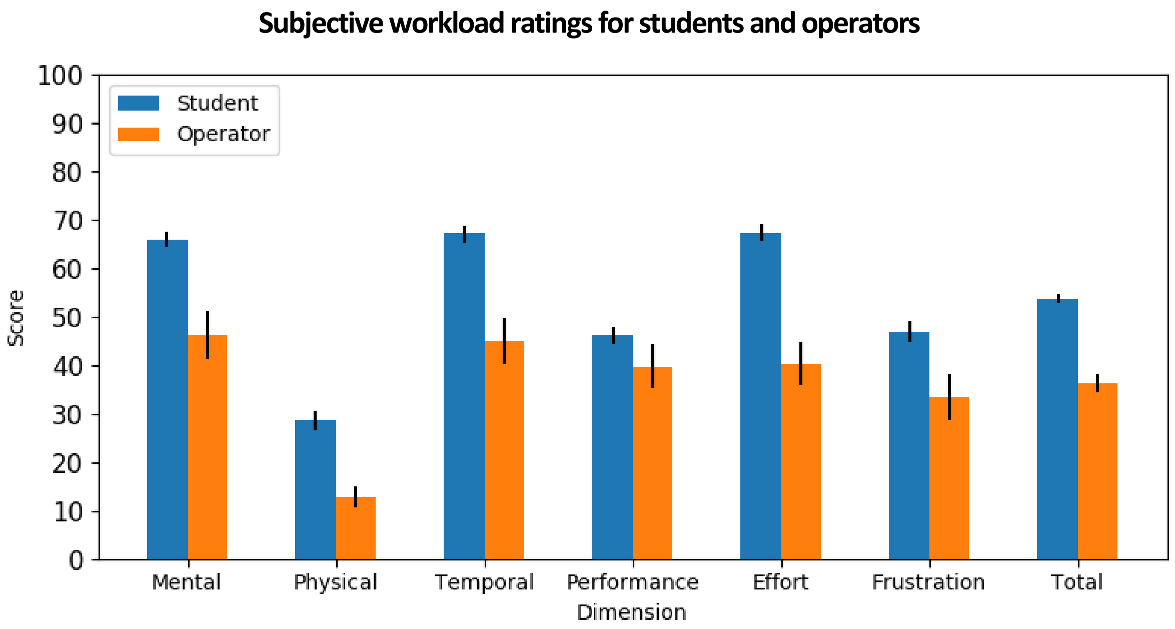


Figure 6.38 A comparison of the subjective ratings of workload reported for each dimension of the NASA-TLX showing the main effect for expertise. Error bars indicate standard error for each mean.

Performance. The same set of performance analyses used in the initial study were also used in this operator expertise study. The total revenue, the time to complete activities associated with modes of operation, and the error in maintaining specific key components at their optimal values were examined across the four experimental trials. With increased exposure and practice monitoring and controlling the microworld, operators were expected to demonstrate increased performance on subsequent trials. Trial performance as measured by the total revenue earned was analyzed across the four trials with a repeated measures ANOVA. A significant main effect for trial on revenue earned was observed, *F*(3,21) = 4.11, *p* = 0.02. As can be seen in Figure 6.39, the operators exhibited increased revenue earned across trials in a similar but more pronounced pattern as students.

A mixed ANOVA with a within-subjects factor of trial and the between subject factor of expertise (student and operator) was performed to compare student and operator revenue across trials. No significant main effect for expertise was found, *F*(1,32) = 0.00, *p* = 0.99. A significant effect for trial on revenue was found, *F*(3,96) = 7.85, *p* = 0.00, but no significant interaction was found between expertise and trial, *F*(3,96) = 2.31, *p* = 0.08.

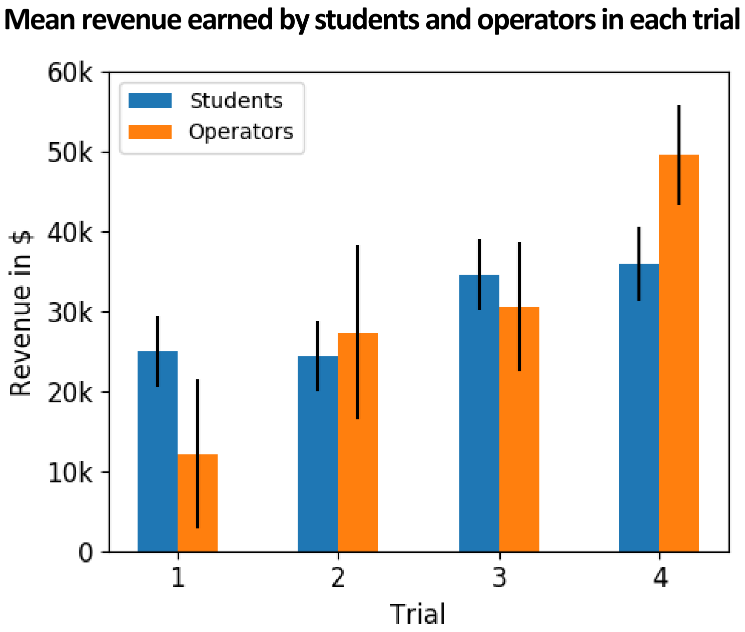


Figure 6.39 A comparison of revenue earned between students and operators across the four experimental trials. Error bars indicate standard error for each mean.

In addition to total revenue earned, the timing measures of performance—including the time to achieve turbine rollup and the elapsed time between turbine rollup and the online mode of operation—were examined across the four trials with two separate repeated measures ANOVAs. A significant main effect for trial on time to achieve the turbine rollup mode of operation was observed for operators, *F*(3,21) = 4.02, *p* = 0.02. Operators became faster at achieving the turbine rollup mode of operation as they completed additional trials, which indicates their understanding and ability to control the simulation improved with experience as predicted.

To further analyze performance in the context of student performance, a mixed ANOVA with a within-subjects factor of trial and the between-subjects factor of expertise was performed to compare student and operator time to achieve the turbine rollup mode of operation across trials. A significant main effect for expertise, *F*(1,32) = 4.45, *p* = 0.04, and trial, *F*(3,96) = 4.22, *p* = 0.01, was observed. No significant interaction was found between expertise and trial, *F*(3,96) = 0.84, *p* = 0.49. A similar pattern of performance as what was observed for revenue is apparent in the time to achieve the turbine rollup mode in which operators performed poorly on the first trial, as shown in Figure 6.40. Operators initially performed more poorly than students, but then improved to the point where they were performing equally well with students in terms of the time to achieve the turbine rollup mode. With additional trials, operators may have demonstrated better performance than students, assuming the trend in the timing data continued.

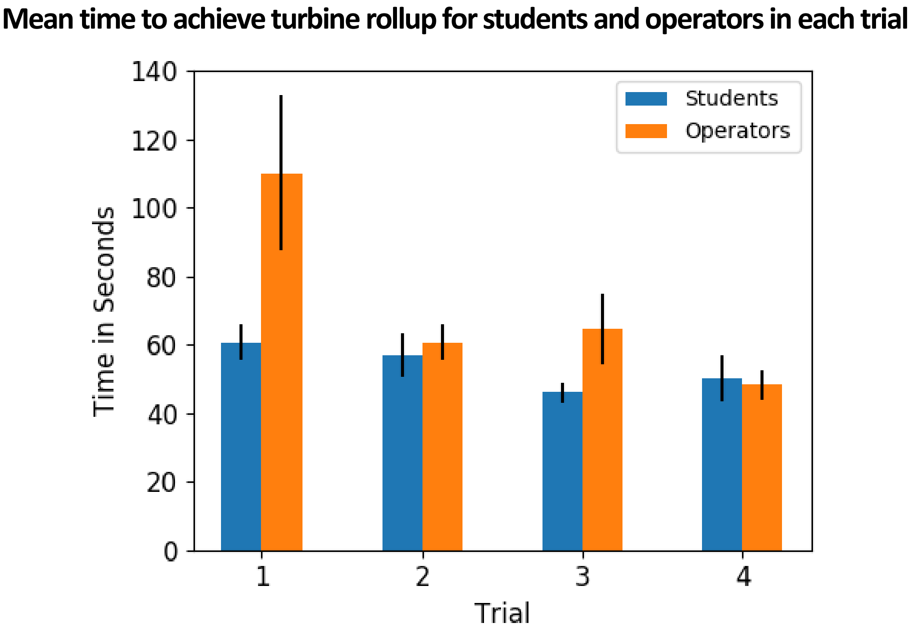


Figure 6.40 A comparison of time to achieve the turbine rollup mode of operation between students and operators across the four experimental trials. Error bars indicate standard error for each mean.

As operators gained experience, the time to transition from the turbine rollup mode to the online mode was predicted to decrease across trials. The time to achieve the online mode of operation was analyzed across the four trials with a repeated measures ANOVA. No significant main effect for trial on time to achieve the online mode of operation, *F*(3,21) = 0.73, *p* = 0.55, was observed for operators when examined alone. Operators did not exhibit shorter times to transition from the turbine rollup mode to the online mode.

Operators were predicted to demonstrate shorter durations between the turbine rollup and online modes as opposed to students, which is indicative of more efficiency and subsequently better performance. A mixed ANOVA with a within-subjects factor of trial and the between subjects factor of expertise was performed to compare student and operator time to achieve the online mode of operation across trials. No significant main effect for expertise, *F*(1,32) = 1.68, *p* = 0.20, or trial was found, *F*(3,96) = 1.28, *p* = 0.28. No significant interaction was found between expertise and trial, *F*(3,96) = 1.21, *p* = 0.31. As can be seen in Figure 6.41, both students and operators spent similar times to transition the simulated plant from the turbine rollup mode to the online mode. This is in contrast to both of the hypotheses that operators would be faster and the previous results for the time to achieve the turbine rollup mode in which operators were faster than students. This transition requires the turbine to move from stopped to an 1800 rpm state, which is limited by a set ramp rate. As a result, the differences in timing achieved by greater expediency may simply have been washed out by the set ramp rate and the associated time the turbine requires to speed up to its online state.

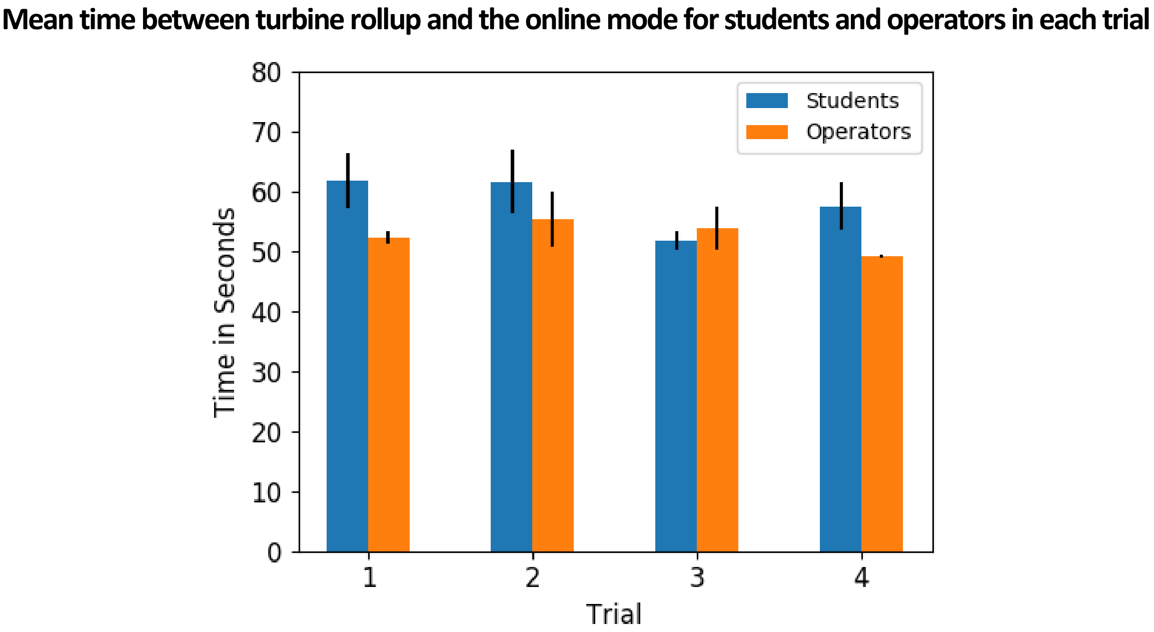


Figure 6.41 A comparison of the elapsed time spanning the rollup mode of operation and achieving the online mode of operation between students and operators across the four experimental trials. Error bars indicate standard error for each mean.

The component-specific measures of performance were also examined. Trial performance as measured by average reactivity was analyzed across the four trials with a repeated measures ANOVA. A significant main effect for trial on average reactivity was observed, *F*(3,21) = 12.85, *p* = 0.00. As can be seen in Figure 6.42, the operators exhibit increased average reactivity across trials in a similar pattern as students exhibited during the initial study.

A mixed ANOVA with a within-subjects factor of trial and the between-subjects factor of expertise was performed to compare student and operator average reactivity across trials. No significant main effect for expertise was found, *F*(1,32) = 0.01, *p* = 0.92. A significant effect for trial was found, *F*(3,96) = 11.08, *p* = 0.00, but no significant interaction was found between expertise and trial, *F*(3,96) = 0.13, *p* = 0.94. In a similar pattern to the other performance measures, operators demonstrated poorer initial performance reflected by a lower average reactivity during the first trial.

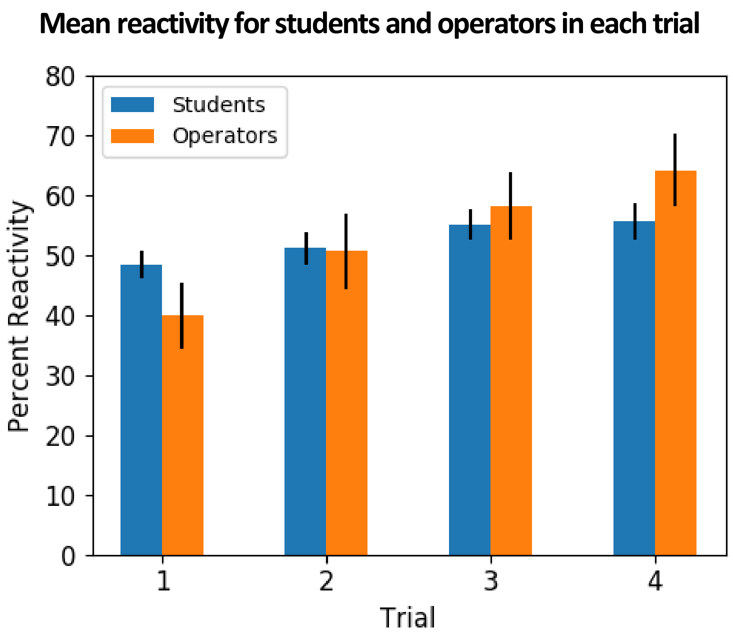


Figure 6.42 A comparison of the average reactivity between students and operators across the four experimental trials. Error bars indicate standard error for each mean.

Trial performance as measured by the average steam generator level error was analyzed across the four trials with a repeated measures ANOVA. No significant main effect for trial on average steam generator level error was observed, *F*(3,21) = 1.105, *p* = 0.37. As can be seen in Figure 6.43, the operators exhibited equivalent average steam generator level error across trials in a similar pattern as students. A mixed ANOVA with a within-subjects factor of trial and the between subject factor of expertise was performed to compare student and operator average steam generator level error across trials. No significant main effect for expertise, *F*(1,32) = 0.10, *p* = 0.76, or trial, *F*(3,96) = 2.72, *p* = 0.06, was found. No significant interaction was found between expertise and trial, *F*(3,96) = 1.23, *p* = 0.30.

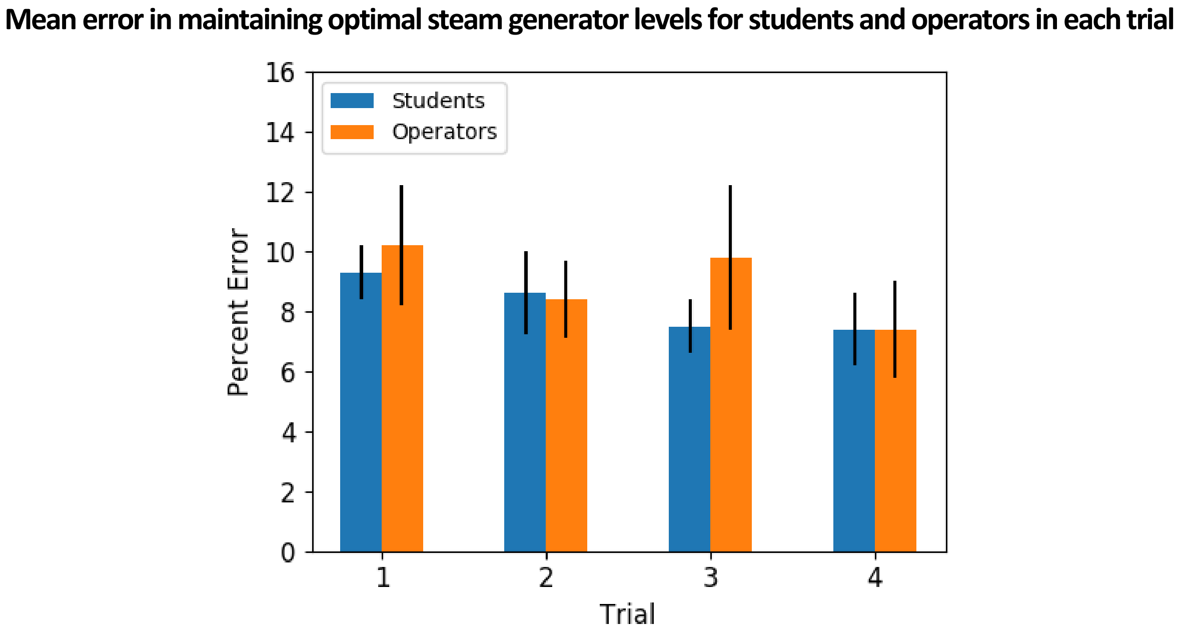


Figure 6.43 A comparison of the average error, defined as deviations from the optimal 50% level, for maintaining the steam generators between students and operators across the four experimental trials. Error bars indicate standard error for each mean.

Recovery time from the two scripted faults was examined for the operators. As with the students, the operators were predicted to demonstrate longer recovery times for the turbine fault than the reactor fault type. This increased recovery time was predicted to translate into decreased total revenue generated during turbine fault condition trials. Two paired-sample t-tests were performed to analyze the time to recover and total revenue earned between the reactor and turbine fault. As predicted, there was a significant difference in the time to recover between the two conditions, $t(7) = 9.21$, $p = 0.02$, with longer recovery times for the turbine fault than the reactor fault as shown in Figure 6.44. A mixed ANOVA with a within-subjects factor of fault type (reactor and turbine) and the between subject factor of expertise (student and operator) was performed to compare student and operator recovery times. A significant main effect for fault type was found, *F*(1,31) = 34.38, *p* = 0.00, in which recovery times for the turbine fault were significantly longer for both students and operators. No significant main effect for expertise, *F*(1,30) = 0.03, *p* = 0.87, or significant interaction between fault type and expertise, *F*(1,31) = 3.87, *p* = 0.06, was found. The lack of main effects for expertise could be due to the small sample size and resulting low-power, since only eight operators were included in the sample. However, the data doesn't reflect a large difference between the two groups, which indicates that there is likely not a significant difference due to expertise.

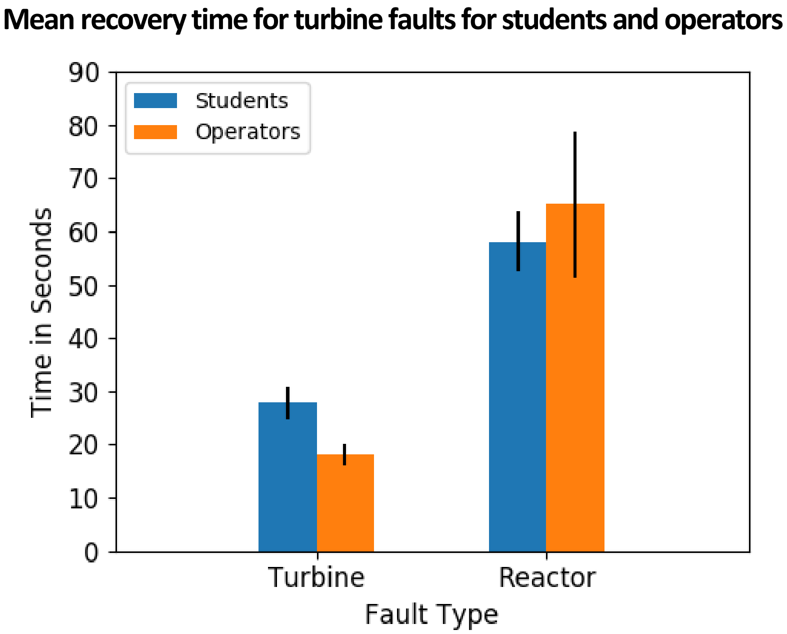


Figure 6.44 A comparison of the recovery times for reactor and turbine fault between students and operators. Error bars indicate standard error for each mean.

Overall, operators exhibited similar pattern of performance to that of students as they interacted with the simulation. Like students, operators demonstrated improved performance on subsequent trials as evidenced by increased revenue and average reactivity. Unlike the students, the operators also demonstrated significant improvement on the time to achieve the turbine rollup mode of operation. The most notable difference between students and operators was found during the initial trial, in which operators consistently demonstrated poorer performance on the different measures. Operators demonstrated lower initial revenue, but then subsequently improved their performances to match that of the students on subsequent trials. This same trend was also apparent in the average reactivity measure in which operators went from initially showing lower performance to matching student performance by the final trial. The initial operator poorer performance followed by a sharp increase that equaled that of the students potentially reflects the differences expertise can impart on a participant’s ability to monitor and control the simulation. Initially, operators struggle because the interface portrays a different system than what they are accustomed to, and as a result, they performed poorly. Once they have a chance to readjust their mental representation of the system, they begin to perform at equal levels as that of the students.

The initial poorer performance demonstrated by the operators could be attributed to negative expertise transfer in which an individual with expertise is hindered by that expertise in a novel situation or domain (Wiley, 1998). The transfer and magnitude of expertise to a new situation or domain is governed by the similarity between the two situations (Kimball & Holyoak, 2000). The operators attempted to draw on their expertise directly, but since the microworld has an objectively different structure from their plant, they overgeneralize relationships that exist within their plant to the simulation and these relationships prove faulty. Once operators were able to determine the distinctions between the microworld and their plant, they could then apply aspects of their expertise, such as how interconnected components respond to each other more appropriately, and in turn, they were able to outperform the student participants.

Situation Awareness. The expert operators have prior experience working in similar contexts at the university steam plant, and were therefore, predicted to exhibit better SA over the novice students. Additionally, the same pattern of results for the freeze probe period and region were also predicted for the operators. Specifically, the operators would demonstrate the most error in the SAGAT-like component value estimates for probes concerning the turbine region across periods, and the second period would demonstrate the highest errors over the other two periods.

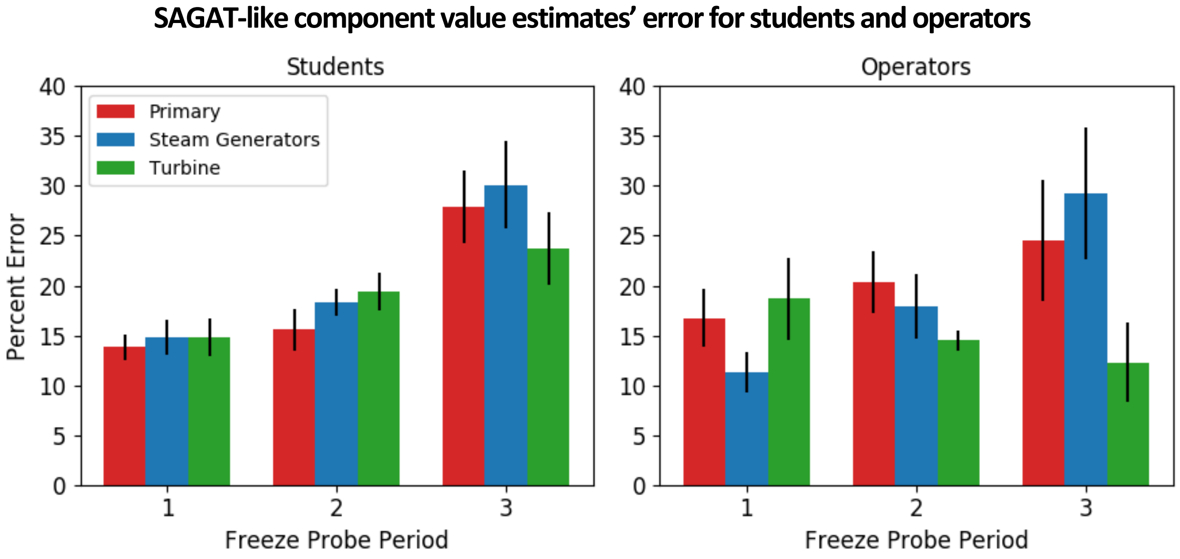


Figure 6.45 Comparisons of the component value estimate response error between students and operators within each region for each of the three sets of freeze probes administered during each trial. Lower percentage error represents more accurate component value estimates and indicates better SA. Error bars indicate standard error for each mean.

A 2 (expertise) by 3 (period) by 3 (region) mixed ANOVA was performed to compare student and operator SA measured by the SAGAT-like component value estimates' error. No significant main effect was observed for expertise, *F*(1,30) = 0.21, *p* = 0.65. The operators did not exhibit lower error in component value estimates as predicted. A significant main effect for region was found, *F*(2,60) = 27.85, *p* = 0.0. As predicted the component estimate errors were the lowest for the primary region and the highest for the turbine region. No significant main effect for period was found, *F*(2,60) = 1.06, *p* = 0.35, but a significant interaction was found between region and expertise, *F*(4,120) = 14.27, *p* = 0.00. As can be seen in Figure 6.45, the interaction between region and expertise results from operators demonstrating significantly lower component value estimate error in the turbine regions for periods two and three than the students. This improved SA exhibited by operators suggests they are better able to respond to probes concerning the most complicated region of the interface than the students. This elevated SA may result from the additional experience and expertise the operators possessed, and therefore, they could better understand the more complicated interrelationships that govern the turbine region components and more effectively estimate the turbine region component values.

Operators SA responses for the SACRI-like trend identification probes were also examined between operators and students. As with the SAGAT-like component value estimate probe types, the operators were expected to demonstrate better SA consisting of more accurate trend identification. Again, operators were predicted to demonstrate the best trend identification accuracy in the primary region and the worst in the turbine region. The operators were also predicted to exhibit the worst trend identification accuracy in the second period when the system is most in flux rather than the first or third periods. A 2 (expertise) by 3 (period) by 3 (region) mixed ANOVA was performed to compare student and operator SA measured by the SACRI-like trend identification probes.

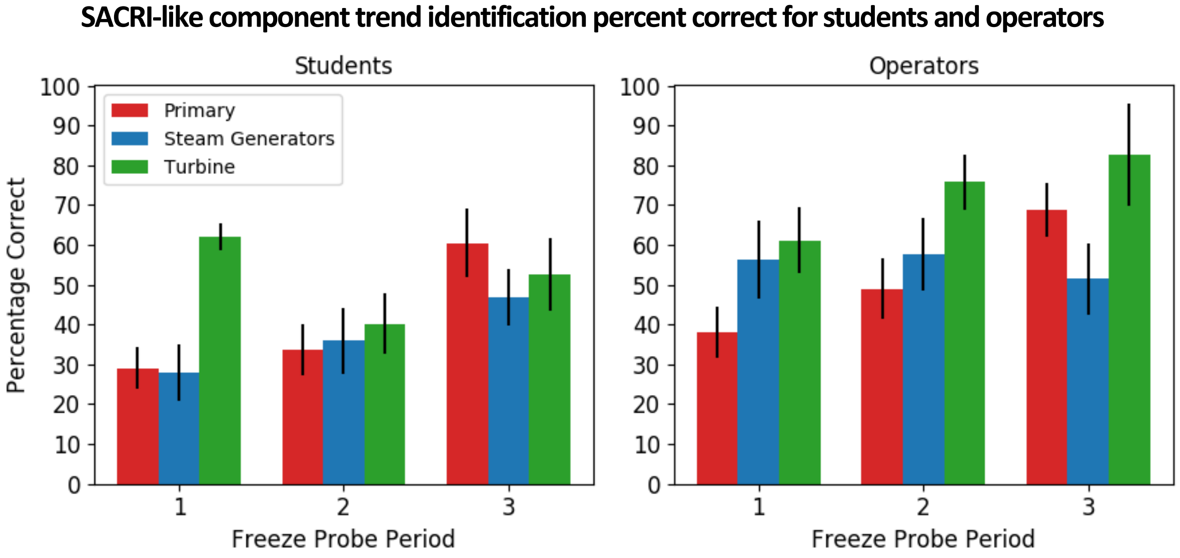


Figure 6.46 Comparisons of the trend identification response accuracy between students and operators within each region for each of the three sets of freeze probes administered during each trial. Higher percentage correct represents more accurate identification of trends and indicates better SA. Error bars indicate standard error for each mean.

A significant main effect was observed for expertise, *F*(1,30) = 6.11, *p* = 0.02. The operators consistently showed better trend identification than students as predicted, as can be seen in Figure 6.46. This significant effect for expertise should be emphasized, since the SA measures previously performed poorly within the student population. The ability to differentiate between novices and students is a substantial finding that provides evidence for the validity of these SA measures. A main effect for region was found, *F*(2,60) = 4.11, *p* = 0.02, and a significant main effect for period was found, *F*(2,60) = 10.50, *p* = 0.00. Students and operators demonstrated higher SA as evidenced by more accurate trend identifications on the first and third periods within the simulation. A significant interaction between region and period, *F*(2,60) = 5.24, *p* = 0.00, and between expertise, region, and period *F*(2,60) = 5.24, *p* = 0.00, was found. These interactions result from students demonstrating equivalent trend identification accuracy for the primary and steam generator regions, while operators showed greater accuracy in the steam generator than primary regions. Both groups demonstrated the highest trend identification accuracy within the turbine region. The operators demonstrated the highest trend identification accuracy for the turbine region regardless of period, while students could only report the turbine trends with high-accuracy while it was not in use, and therefore, unchanging. Operators were able to maintain a high-level of accuracy for turbine trend identification at periods two and three as well. The turbine region is the most complicated region of the interface when it is in operation and receiving steam. The operators were able to effectively report the trend direction of turbine region components with considerable accuracy while it was in operation, during periods one and two. The pattern of attention exhibited by operators, which is described in the next analysis section in detail, does not reflect additional attention towards the turbine region that could account for better knowledge of its state. Furthermore, the patterns of attention between students and operators did not differ much. As a result, operators appear to have a better mental representation that allows them to more effectively track and integrate the various component values to allow them to more efficiently acquire information necessary for more accurate trend identifications.

Attention. As with the students and following the same analyses approach, the patterns of attention defined as the proportion of attention allocated to each of the four regions of the interface were examined during each of the activities associated with the startup, rollup, and online modes of operation within each of the four experimental trials. The same patterns of attention demonstrated by the students were predicted for the operators. Specifically, the patterns of attention were predicted to vary based on the current process control activity. During the startup mode of operation, a higher proportion of attention was predicted to be allocated to the primary region, since this region of the interface contains the indicators and controls to support controlling reactivity and heating the reactor coolant. During the rollup and online modes of operation, the distribution of attention was predicted to be more evenly distributed across the primary, steam generators, and turbine regions. The alarm region was predicted to receive less attention than the other three regions during all modes of operation, since it provides redundant information and is of little extra value for monitoring and controlling the simulation. Separate 2 (measurement type) by 3 (mode) repeated measures ANOVA were performed to analyze the proportions of attention within each region across the modes of operation. The results of these analysis can be seen in Table 6.13 and a graphical representation of the main effects and interactions can be seen in Figure 6.47.

Table 6.13 Results from the ANOVA analyses of the two measurement types, acknowledgement and fixation proportions across the three modes of operation for the expert operators.

	Source	df	df_{error}	F	p	η^2	*Power*
Alarms	Type	1	6	0.22	0.66	0.03	0.07
	Mode	2	12	1.84	0.20	0.23	0.31
	Type * Mode	2	12	0.44	0.65	0.07	0.11
Primary	Type	1	6	0.64	0.46	0.10	0.10
	Mode	2	12	27.53	0.00*	0.82	1.00
	Type * Mode	2	12	2.32	0.14	0.28	0.38
Steam Generators	Type	1	6	0.08	0.78	0.01	0.06
	Mode	2	12	18.29	0.00*	0.75	1.00
	Type * Mode	2	12	5.08	0.03	0.46	0.71
Turbine	Type	1	6	3.34	0.12	0.36	0.34
	Mode	2	12	10.04	0.00*	0.63	0.95
	Type * Mode	2	12	0.15	0.86	0.02	0.07

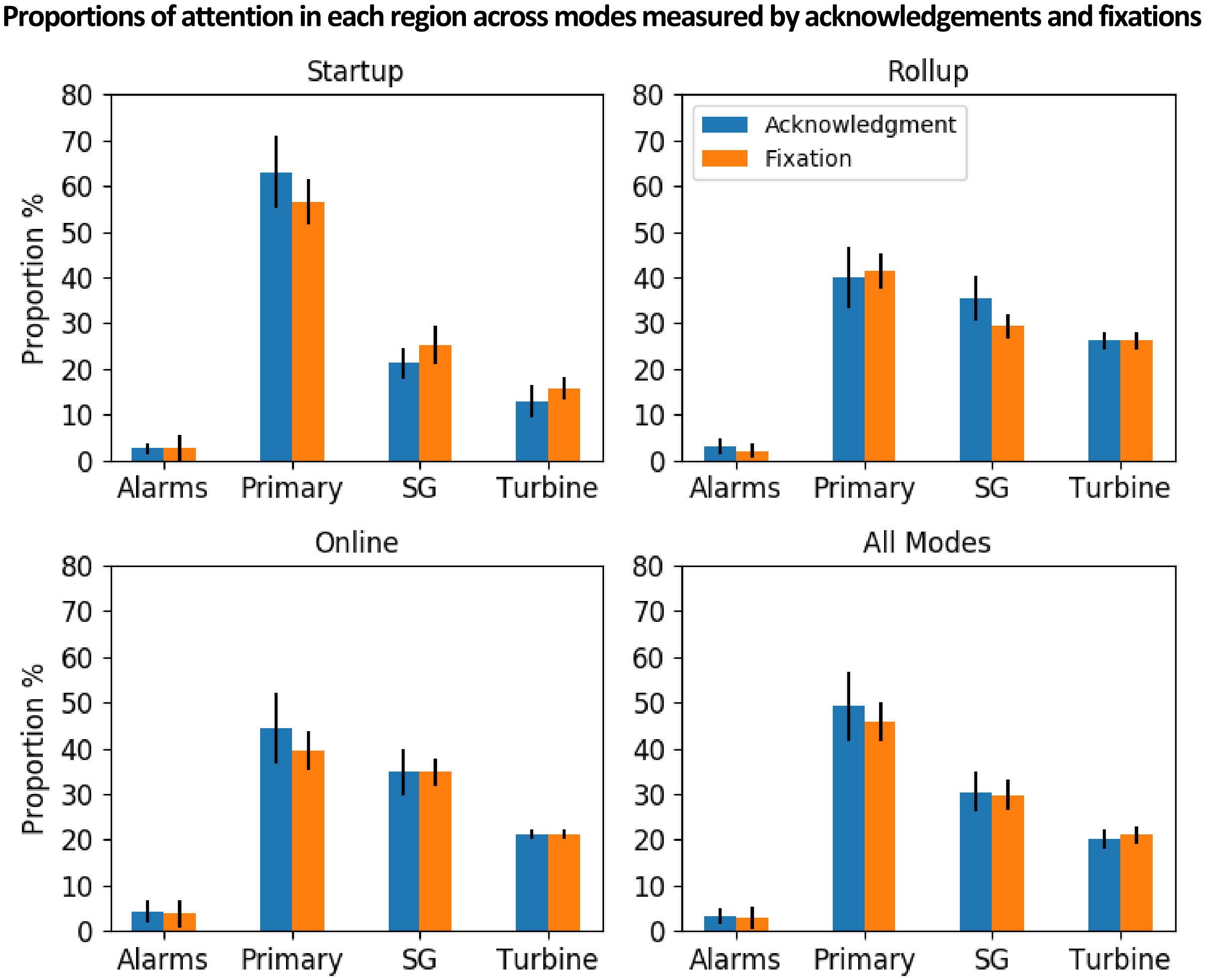


Figure 6.47 A comparison of the proportions of attention within each region for each mode of operation to show the main effects and interactions for measurement type and mode of operation. Error bars indicate standard error for each mean.

One of the central research questions concerns whether a student population with little training could serve in place of more-difficult-to-obtain expert operators. To determine the representativeness of a student population that could be used for research in place of expert operators, following the concept of discount usability (Nielson, 1989), it is important to compare the patterns of attention exhibited by students and operators. By comparing the patterns of attention, it is possible to determine if the attentional component of SA is similar between students and operators. Additionally, it is also of interest how the acknowledgement and fixation measures of attention compare to further validate the use of the acknowledgement measure in lieu of eye tracking. The operators were predicted to demonstrate a more optimal pattern of attention due to their expertise. To make the comparison between experts the patterns of attention within regions of the interface were analyzed for the fixation and acknowledgement measures of attention. To analyze the patterns of attention measured by fixations, separate 2 (expertise) by 3 (mode) mixed design ANOVAs were used. The results of these analyses are summarized in Table 6.14.

Table 6.14 Results from the ANOVA analyses of students and operators for the proportion of fixations within each region across modes.

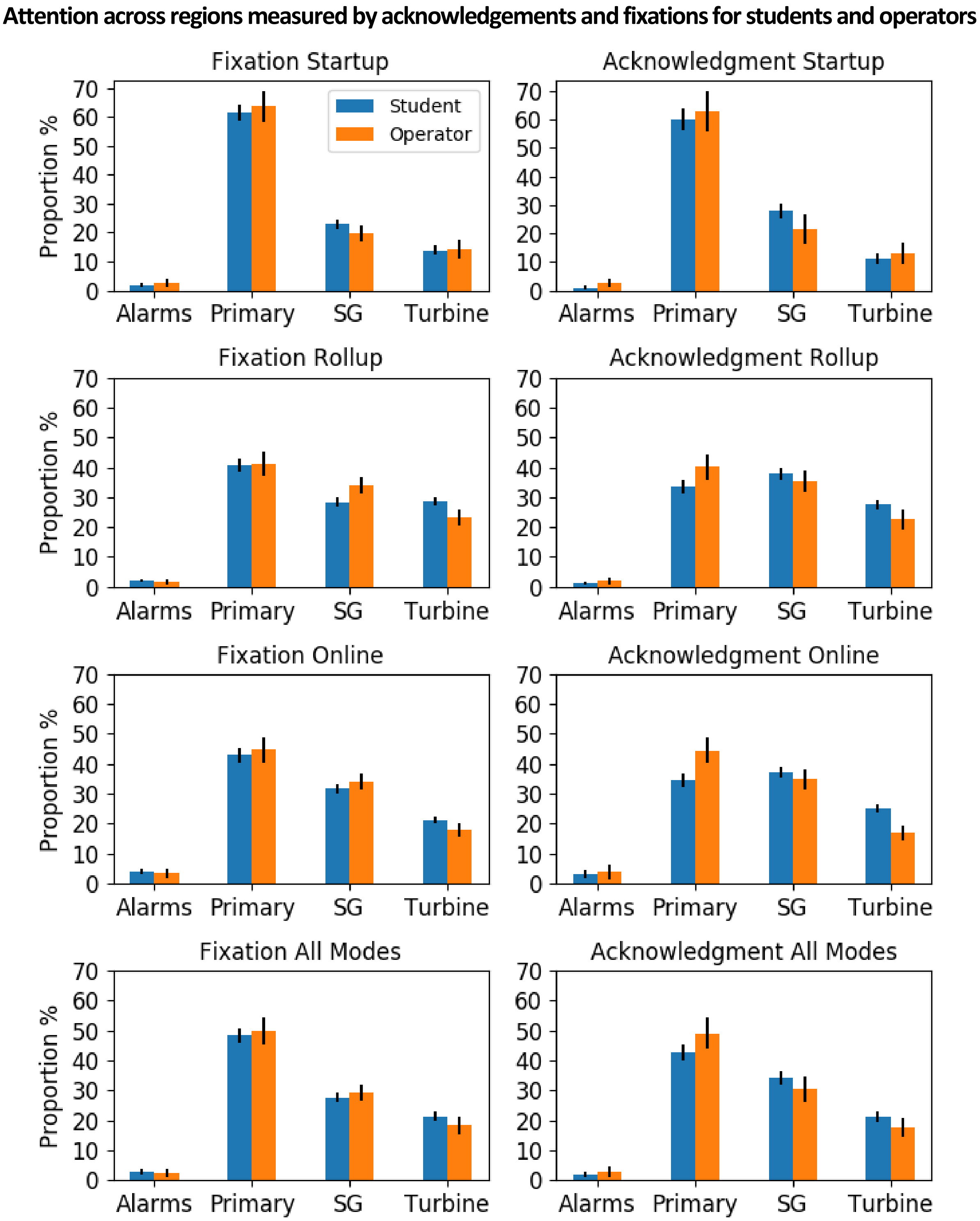
	Source	**df**	**df$_{error}$**	***F***	***p***	**η^2**	***Power***
Alarms	Mode	2	64	5.26	0.01*	0.14	0.82
	Group	1	32	0.02	0.89	0.00	0.05
	Mode * Group	2	64	0.67	0.52	0.02	0.16
Primary	Mode	2	64	54.15	0.00*	0.63	1.00
	Group	1	32	0.09	0.76	0.00	0.06
	Mode * Group	2	64	0.11	0.90	0.00	0.07
Steam Generators	Mode	2	64	42.83	0.00*	0.57	1.00
	Group	1	32	0.41	0.53	0.01	0.10
	Mode * Group	2	64	5.15	0.01	0.14	0.81
Turbine	Mode	2	64	28.69	0.00*	0.47	1.00
	Group	1	32	1.18	0.29	0.04	0.18
	Mode * Group	2	64	1.61	0.21	0.05	0.33

Another set of analyses using Separate 2 (expertise) by 3 (mode) mixed design ANOVAs were performed to examine differences in patterns of attention measured by the acknowledgement measure between the students and the operators. The results of these analyses are summarized in Table 6.15.

Table 6.15 Results from the ANOVA analyses of students and operators for the proportion of acknowledgements within each region across modes.

	Source	df	df_{error}	*F*	*p*	η^2	*Power*
Alarms	Mode	2	64	2.52	0.09	0.08	0.49
	Group	1	32	0.42	0.52	0.01	0.10
	Mode * Group	2	64	0.31	0.73	0.01	0.10
Primary	Mode	2	64	34.51	0.00*	0.53	1.00
	Group	1	32	1.82	0.19	0.06	0.26
	Mode * Group	2	64	0.62	0.54	0.02	0.15
Steam Generators	Mode	2	64	16.94	0.00*	0.35	1.00
	Group	1	32	0.97	0.33	0.03	0.16
	Mode * Group	2	64	0.52	0.60	0.02	0.13
Turbine	Mode	2	64	21.03	0.00*	0.40	1.00
	Group	1	32	1.79	0.19	0.05	0.25
	Mode * Group	2	64	3.09	0.05	0.09	0.58

Main effects for mode were found consistently by both measures of attention. These significant effects for mode demonstrate the effectiveness of the proportions of fixations and acknowledgements to characterize patterns of attention based on specific task demands. There was no significant effect for expertise on the proportion of attention allocated to each region as measured by either fixations or acknowledgements, as seen previously in Table 6.14 and Table 6.15. Operators did not demonstrate significantly different proportions of attention to any of the interface regions. It would be inappropriate to conclude that operators are equivalent to students based on interpreting the lack of a significant main effect, since the lack of significance does not mean an actual difference is not present. Rather it may suggest the experimental design and sensitivity of the acknowledgement and eye tracking fixation measures to detect differences are not sufficient. However, qualitatively the patterns of attention for both students and operators are quite similar, as is shown in Figure 6.48. This similarity, taken together with the lack of statistically significant main effects for any region, suggest it is reasonable to conclude that operators and students do not differ in their patterns of attention in a meaningful way. A significant main effect for type of attention measurement was found for the primary and steam generator regions. The fixation measure captured a higher proportion of attention to the primary region and a lower proportion of attention to the steam generator region. As mentioned for expertise, the overall pattern for both the acknowledgement and fixation measures collectively across all regions are qualitatively quite similar, and therefore, there is evidence to suggest the two measures of attention are in fact capturing attention in similar manners.



Attention across regions measured by acknowledgements and fixations for students and operators

Fixation Startup

Acknowledgment Startup

Fixation Rollup

Acknowledgment Rollup

Fixation Online

Acknowledgment Online

Fixation All Modes

Acknowledgment All Modes

Figure 6.48 A comparison of the proportions of attention within each mode for students and operators. Error bars indicate standard error for each mean.

Mental Model. The mental model measure demonstrated promise in the initial study concerning students. The mental model measure correlated well with performance, and therefore, it was also included as a measure for the operators. The component and indicator completeness scores were correlated with performance, SAGAT-like SA, and SACRI-like SA measures, as can be seen in Table 6.16.

Table 6.16 Correlations between the mental model component and indicator measures and performance, SAGAT-like value estimate error SA, and SACRI-like trend identification accuracy SA measures. No significant correlations were found.

Mental Model Score	Measure	*r*	*p*
Component			
	Revenue	0.45	0.27
	SA Error	0.56	0.15
	SA Trend	0.19	0.66
Indicator			
	Revenue	-0.05	0.91
	SA Error	-0.69	0.06
	SA Trend	0.24	0.57

The mental model measure did not perform as well in the operator population. There were not significant correlations between either mental model measure, or any of the other measures. Next, the mental model component and indicator scores were compared between the students and the operators. No significant difference was found between the students and operators for either the component completeness score, *t*(31) = -1.39, *p* = 0.18, or the indicator completeness score, *t*(31) = 0.03, *p* = 0.98. Based on this finding, the operators and the students did not exhibit much difference in their mental model of the system. This supports the notion that the expertise operators used to achieve superior SA as measured by the SACRI-like trend deification accuracy measure does not rely on a more accurate mental representation of the system, but rather a better ability to track the fluctuation of values and their inter-relationships. This demonstrates a clear distinction between students and operators in terms of their ability to perform process control tasks.

Chapter 7. Discussion

There were three main objectives for this project. The first centered on developing a microworld platform that could be used for human factors-based psychological research. The second focused on developing and evaluating a novel measure of attention to aid in the evaluation of SA and assess attention patterns within the context of human performance. The third dealt with comparing a novice student population with experienced operators to determine whether microworld-based studies can be leveraged to examine human performance in an easier-to-sample population. The microworld simulation, coined “Rancor,” was developed to serve as a platform for conducting research with student populations to extend the ability of researchers to conduct complex process control outside of applied industrial settings (Ulrich, Werner, & Boring, 2015). The Rancor Microworld demonstrated promise as an experimental platform for human factors research, which is an important outcome for this project since this platform can be used in conjunction with full-scope simulations conducted for control room modernization efforts. The proposed attention-acknowledgement measure demonstrated promise as a means to simply assess attention. The attention-acknowledgement measure can be used in full-scope simulation studies as part of control room modernization, but to do so would require additional development to scale the measure from a single display used in the Rancor Microworld to the full-scope simulator with its many displays. Lastly, the comparison objective of the novice students and experienced operators revealed some important distinctions between the two groups, specifically in regard to SA, but there is still ample opportunity to use students in place of operators for other evaluations such as attention and potentially even performance. As a platform to evaluate novel designs, the Rancor Microworld provides promise for using students to vet designs prior to their evaluation with actual operators on the full-scope simulator used during control room modernization studies. The following sections describe the key findings from this series of studies and describe the importance of these findings.

Rancor Microworld

The results from the three studies demonstrate the efficacy of a nuclear process control microworld to conduct research on SA. The microworld simulation demonstrated promise as a research platform in which students and expert operators were able to rapidly learn to control effectively. The microworld contains sufficient complexity to generate meaningful and significant effects for variables of interest, such as the performance measures. Total revenue as a performance measure demonstrated significant effects for learning as evidenced in the improvement over subsequent trials for students in the initial study and marked improvement exhibited by participants who returned for the follow-up experiment session. Total revenue was the least sensitive of the measures, while more specific measures, such as reactivity, were useful for examining differences in control strategies. The variability in the total revenue could be explained in large part by a control strategy in which participants who operated the reactor at a higher level of reactivity earned more revenue than participants who followed a more conservative lower level of reactivity, which reduced the dynamics of the system and made the overall process simpler to control.

The microworld also generated significant effects for cognitive factors as illustrated by the NASA-TLX subjective workload scores provided by participants after each trial. The NASA-TLX was capable of detecting differences between students and operators, as well as differentiating students in the initial and follow-up studies, based on their reported workload. Furthermore, the levels of workload exhibited by both students and operators were moderate in comparison to other studies in hospital and surface transportation domains (Hoonakker et al., 2011; Lee et al., 2001), which suggests the microworld is not overly difficult nor is it trivial to control. Therefore, the microworld shows promise as a platform for examining psychological constructs within the context of complex process control. The Rancor Microworld was also capable of generating significant differences in SA as measured by the SAGAT-like component estimate error and SACRI-like trend identification accuracy measures included in this series of studies.

Operators versus students

One of the key components of this research concerned comparing operators and students to determine the suitability of student participants to conduct inexpensive process control-based research that is generalizable to expert populations. Operators demonstrated some meaningful differences that raise some potential concerns for using student populations as a substitute, but they also demonstrated a large degree of similarity.

The students and operators differed in the subjective mental workload they reported. Operators reported lower mental workload. This lower workload did not necessarily translate into better performance, however. The operators did not generate significantly more revenue. One notable difference between the two populations is the trend in the performance measures in which the operators exhibited poorer performance than the students on the first trial. Care must be taken to interpret this finding, since this trend only demonstrated a statistically significant difference on one of the performance measures, which was the time to achieve the turbine latch mode of operation. Several of the other measures, which did not demonstrate a statistically significant difference, did qualitatively show that the operators initially performed more poorly than the students. Only eight operators could be included in the expert sample, and due to this sample size and the resulting reduction in power, it was difficult to find a significant effect. The total revenue earned by the operators was lower than that of the students on the first trial, but then reached levels equivalent to that of the students by the fourth trial. The reactivity performance measure showed the operators initially exhibited lower levels of reactivity and then matched that of the students on subsequent trials. The operators had more difficulty initially because of negative expertise transfer (Wiley, 1998), which interferes with an accurate conceptualization of the process control simulation. Operators likely drew on assumptions from their steam plant operations that did not hold true in the microworld, and therefore, they impeded the ability of the operators to effectively control the simulation during the first trial. Once the operators became accustomed to the specifics of the system and recalibrated their assumptions (Hsu, 2006), they were able to reach levels of performance equivalent to that of the students.

The timing data provided mixed results in which the operators were slower to achieve the turbine rollup mode, but they exhibited similar times to that of the students for the elapsed time between the turbine rollup and achieving the online mode of operation and the recovery time for the scripted turbine fault. Since the timing measures for the transition between the turbine rollup and online mode and the recovery period both require the turbine to be ramped up to speed at a set rate over time, it is likely that any effects of expertise and the associated efficiencies that reduce the required time were washed out by the fixed ramp rate of the turbine. The most diagnostic of the three timing measures was the time to achieve the turbine rollup. This time is governed by the system's time dynamics, but these time dynamics are based primarily on the configuration of the system, via the speed at which participants add heat by increasing reactivity. Thus, the time to achieve turbine rollup is more responsive to participant actions. The time to achieve turbine rollup is a better measure because the difference between students and operators was detected with the operators demonstrating slower times to achieve turbine rollup. In another performance measure, students and operators did not differ in their ability to control the steam generator level as evidenced by their similar error percentages. Since the reactivity performance measure, which also concerned the participants' control strategies, did find a significant effect, it was surprising that no significant effect for expertise was found. The steam generators fluctuate slowly, which simply may have reduced the challenge of maintaining their levels near the optimal 50% value. As a result, the students and operators were both able to control these with low-errors, which may have prevented a significant main effect from being found.

The students and operators did not differ in the patterns of attention they exhibited. Of the four analyses performed to compare the proportions of attention within each region and between modes of operation, only a single significant main effect for expertise emerged. Though a significant effect was revealed, no family-wise error corrections were used in these analyses, and therefore, the single significant effect could simply be the result of chance. The effect size for expertise in each region was also quite small, ranging from 0.00 to 0.02, which indicates any true difference is minor, if in fact it did exist. Furthermore, from a qualitative perspective, the proportions of attention follow similar trends and do not appear to differ much. Collectively, the results provide support that students and operators do not differ in the overall patterns of attention they exhibit as they interact with the simulation. The patterns of attention captured by the two measures reflect a typical pattern of attention dictated by the task, as would be expected by the SEEV model (Wickens et al., 2003). This is not to say that student and operator attention is similar in all respects, such as their eye-scanning patterns of order of fixations and dwell-times (Kasarskis et al., 2001). However, in terms of performing a task within the microworld, the general proportions of attention allocated to various regions are analogous between students and operators. Since participants and students behave similarly, attention is one area in which students may serve as a good alternative to operators. For example, within the context of control room modernization, students could be used to vet the usefulness of a new interface design that incorporates a new visual design element by capturing the proportion of attention they allocated to the new design element in relation to a system without the new design element. The evidence here suggests this proportion of attention should generalize to experts, and therefore, provide an indication of how useful this new design element will be when used by experts.

There were also disparities between operators and students in regard to the SA they exhibited. The following section dedicated to SA will include a discussion of the differences between students and operators in the context of the SA measures used in this study.

Situation Awareness

Two different types of freeze probes were used to assess SA within this series of studies. The two measures captured different aspects of users SA concerning the current system state. The SAGAT-like probes consisted of component value estimates that were used to generate errors in the estimates and assess SA from the magnitude of these errors (Endsley, 1995b). The SACRI-like probes consisted of more basic information concerning how the value of a component fluctuates over time as a trend. The SACRI was developed explicitly for process control and has been used by Hogg et al. (1995) in full-scope simulation studies. The SAGAT was developed for use in aviation with more dynamic situations and situation elements that do not readily lend themselves to trends, such as the existence of a nearby aircraft (Endsley, 1995b). The two measures did demonstrate significant correlations, and therefore, are assessing similar aspects of SA. The results from this series of studies provides evidence that the SACRI-like probes are effective in diagnosing differences in SA between novice and expert populations. SA as measured by the SACRI-like probe questions demonstrated a significant difference between students and operators. The analyses of the SAGAT-like probe questions examining the error for component value estimates did not find a main effect for group; however, a significant interaction between group and period was found. This interaction is notable because it illuminates a key difference in SA between students and operators in how they are able to respond to probes concerning regions of differing complexity. The probes concerning the turbine region proved the most challenging for students. The three periods were progressively more challenging for students as the trial unfolded with the error in estimated component values increasing with each subsequent period. In contrast, the operators did not experience any additional difficulty in providing error estimates for component values as the trial progressed with similar estimated value error percentages observed in each time-period that were not significantly different from each other. This demonstrates that the operators were not susceptible to the increased difficulty for each period as was the case for the students.

To the best of the author's knowledge, this is the first instance of using a SACRI-like freeze probe in a microworld setting. One beneficial outcome of this research is the successful demonstration of this particular freeze probe technique in a microworld setting over short-duration tasks. The SAGAT-like probes were capable of consistently diagnosing difference in SA based on the region of the interface in each of the three studies. However, the SAGAT-like probes were less effective than the SACRI-like probes and did not yield a significant main effect for expertise since they could not differentiate students from operators. Both freeze probe questions provide complimentary information and both are easily administered concurrently; therefore, there is a benefit to including both types of probe questions in future research.

It was not initially predicted that either SA measure would fail to show a meaningful relationship to performance. The SACRI-like SA measure was able to differentiate between students and operators; however, overall performance on the measure was much poorer than what was observed by Hogg et al. (1995). The overall low-SA performance observed by students and operators suggests that the SACRI-like measure may not have had sufficient sensitivity to account for the subtle differences in revenue within the initial student participant study, while it was able to distinguish between the student and operator populations. It is quite possible that the SA measure simply did not perform adequately enough to detect these subtle performance differences. The students interacted with the simulation for a brief period of time, spanning less than two hours total. This time was sufficient to allow them to successfully control the simulation, but it may not have been sufficient to allow them to respond to the freeze probes as evidenced by the overall low-SA scores they demonstrated. Even after the additional exposure of another two hours, the students in the follow-up study did not show an increase in SA. Even more experience may be needed to allow novices to understand the system sufficiently so they can report SA to higher levels and avoid a floor-effect for the measure. Additionally, the freeze probes could be refined to a more restrictive list that is more manageable for students. This would sacrifice capturing their knowledge of the whole system, but it could prove more diagnostic to evaluate their SA concerning a more limited but more important section of the process control task. To refine the list, it would be beneficial to recruit a NPP operator, train them on the system, and then subject them to the complete list of freeze probe items at each time-point. This approach would require considerably more time than the protocol used here, which presented participants with randomly selected questions, but it would support an item analysis that could be used to refine the freeze probe measure and make it more diagnostic. Future studies are encouraged to continue these efforts and investigate a more refined set of freeze probes.

The lack of a clear relationship between SA and performance can be explained by other reasons, such as the disjunction between performance and SA, which is a fundamental defining feature of the SA construct (Endsley et al., 1998; Durso et al., 1998). SA is required, but not sufficient, to achieve good performance. It is this disconnect that could account in part for the lack of a clear relationship. Indeed, others have reported a lack of a relationship between SA and performance (Demas, Lau, & Elks, 2015; O'Brien & O'Hare, 2007). Another explanation for the lack of a clear relationship between performance and SA concerns the level of SA assessed with these freeze probes. Endsley's model contains three levels, which are the knowledge of system components, the integration of the components into a cohesive mental representation, and the prediction of future states of the system (Endsley, 1995a). The freeze probes in this current project assessed SA primarily at Level I, but also at Level II since the system is highly coupled and allows for determining the values or trends of a given component based on knowledge of a related component. Since SA was not evaluated at Level III, it is possible that the success or failure of achieving Level III can account for some of the differences between SA and performance. Others have found significant correlations with SA Level II and Level III probes and performance, while Level I probes failed to significantly correlate (O'Brien & O'Hare, 2007). Indeed, the control strategy of operating the plant at a higher reactivity requires participants to project the effects of the higher reactivity to respond to transients in sufficient time, since the plant fluctuates more rapidly when additional heat from greater reactivity is constantly introduced into the system. The lack of a clear relationship between SA and attention will be discussed in a subsequent section devoted to the attention-acknowledgement measure.

Mental Models

The mental models of participants were assessed based on system sketches completed at the end of the experimental sessions in the initial student study and operator expertise study. The mental model measure correlated with performance in the initial study, but it did not correlate with performance in the expertise study comparing operators to students. These mixed results suggest the measure may demonstrate usefulness in explaining human performance. The measure may be effective as a means to explain performance for untrained novice participants, since the student performance was significantly correlated with their mental model scores. Responses to probes concerning the values and trends of components and indicators proved too challenging for student participants, while the mental model measure did not require this specific level of knowledge. The distinction between a mental model and SA merit the use of the mental model measure, though the two terms have been blended together by the research community.

It should be noted that the term “mental model” has been used in the SA literature to refer to the current state of SA at a point in time (Sarter & Woods, 1991). The article this appeared in was criticizing SA for its ill-definition and using the term too generally. Using SA to also encompass the mental model of the underlying system is disadvantageous because it combines the structural component of mental model, which defines the interrelationships of components with the real-time values these components take on while a participant is interacting with the simulation. Certainly, poor performance could result from a failure to acquire the component or indicator value, but poor performance could also result from issues with the mental model based on an inappropriate structure or even the complete exclusion of key elements. This later type of issue is central to the approach followed in this project, since the mental model measure was envisioned to add explanatory power by determining which components and indicators participants included in their system sketch. In this manner, the mental model measure provides an additional dimension to potentially explain performance beyond the SA freeze probe measures’ assessment of the current values and trends of components and indicators. The results also demonstrated the distinction between the mental model and the SA based on the lack of significant correlations between two types of measures. The mental model measure could be improved upon with further refinement, such as directing participants to recreate only the critical aspects of the system. This restriction to critical aspects may provide more insight into the actual mental model as opposed to generic instructions to recreate the system. The mixed success of the mental model measure also merits further research since SA is often not correlated with performance, and therefore, other ways of explaining human performance are needed.

Attention-Acknowledgement Measure

As a method to assess patterns of attention-allocation, the acknowledgement measure was effective and demonstrated several positive characteristics that support its use in future research. In contrast to eye tracking, the acknowledgement measure doesn't suffer from data loss due to the technical challenges associated with capturing the corneal reflection, pupil, and head position. The average data loss for each participant within this project was 6.50% (*SD* = 4.15%). This is on the lower end of the range of data loss reported by others, which can range as high as 60 percent in eye tracking studies (Holmqvist et al., 2011). In regard to assessing the same construct of attention, the measure qualitatively matched the same pattern of attention assessed by the fixation-based measure in each of the modes of operation, and in general, across all modes of operation. The general pattern of attention characterizing participants' interaction with the simulator consisted of approximately 5, 45, 30, and 20% to the alarms, primary, steam generators, and turbine regions of the interface, respectively. The measure was capable of consistently differentiating between patterns of attention for the different activities performed within the analyzed modes in each of the studies. Indeed, significant main effects were found for mode in each study. The alarm region did not show a significant effect for mode, but this was an expected result. The alarms are redundant and consist of largely unused information and served primarily as a control to provide a region in which attention was expected to be allocated in the lowest proportion. This was demonstrated to be the case as the alarms consistently received the lowest proportion of attention. Since participants were predicted to largely ignore the alarms, it is in line with the predicted hypothesis that the mode has no impact on the amount of attention allocated to the alarm region.

The notable exception to the acknowledgement measures' effectiveness is the low-temporal resolution and the resulting inability to capture attention patterns for short-duration tasks. Indeed, the two short-duration modes, ready to roll and ready to sync, did not generate sufficient acknowledgement data to support a comparison against the fixation-based measure. Additionally, since the measure requires an active selection to designate an acknowledgement of a marker, the measure potentially competes with other control actions to further compound the challenge of capturing attention during these short-duration modes. Furthermore, there were significant differences in the proportions of attention measured by acknowledgements and fixations during the turbine recovery event, which required a large number of control actions. These control actions directly compete against performing the acknowledgements, and therefore, the acknowledgement measure did not accurately capture attention as the fixation measure was able to, especially during the periods requiring many control inputs. Over larger timescales, the measure performed quite well, and qualitatively, the pattern of attention measured by the acknowledgements closely resembled the pattern measured by fixations.

The acknowledgement measure did show some slight interference with the primary task; however, the results from the NASA-TLX workload measure overall indicates this interference is minimal. The students in the follow-up study did not report significantly greater workload on the NASA-TLX and the pattern of performance in terms of revenue, reactivity, and steam generator level deviations did not differ significantly.

The general lack of the predicted relationship between the attention measures and the performance likely stems from a small interface size, and the highly coupled nature of the components. Attention becomes more diagnostic when there are limited attentional resources that must be strategically allocated to critical locations and information. For example, in a full-scope simulation, the attentional demands are sufficiently high that operators simply cannot scan the entire interface and process all necessary values, at least within brief timeframes, to diagnose an issue or assess current plant status. Therefore, operators must rely on selective attention to use their limited attentional resources to identify the critical information needed for the task (Broadbent, 1958; Driver, 2001). The small interface for the microworld allows participants to scan the entire interface, even in short durations. Due in part to its small size, which was necessary to maintain a reduced scope, the microworld simulation did not force participants into situations in which they had to forgo checking a value, and therefore, shifts in the proportions of attention between the three regions did not effectively capture any performance differences. In line with the assumptions of the SEEV model (Wickens et al., 2003), the primary region was deemed critical for the process, and indeed, the attention measures exhibited the highest proportions of attention to this region. However, beyond being able to verify the importance of the primary region, the proportion measurements of attention could not account for specific instances of attention, which would require more time-sensitive measures, such as the scan paths. These scan paths were not investigated, since the acknowledgement measure does not have sufficient temporal resolution to support this type of analysis, which is another limitation of the measure.

The same rationale for the lack of a relationship between attention and performance can also be applied to the apparent lack of a meaningful relationship between attention and the SA measures. The highly coupled components and small interface size rendered attention less sensitive and precluded using attention to assist in explaining the differences measured in SA. Though there were no consistent or meaningful relationships between errors or trend SA measures and proportions of attention, the general patterns of attention do appear to match overall SA performance to some degree. Participants dedicated the majority of their attention towards the primary region, a moderate amount to the steam generators region, and the smallest amount to the turbine region. The primary region demonstrated the lowest percentage of error, the steam generator region a higher percentage error, and the turbine region the highest percentage of error in the SAGAT-like component value error SA measure. This pattern of percentage error in the SAGAT-like SA measure increases as the proportion of attention within each region decreases. There should be a relationship between SA and attention and future studies with less coupled component configurations should be performed to identify and confirm this relationship.

As previously noted, the acknowledgement measure did not perform well during the recovery period for the scripted turbine fault due to the substantial number of competing control actions participants were required to perform as part of the primary task. To overcome this disadvantage, the control actions could potentially be combined with acknowledgements to serve as an aggregate measure of attention. Combining control actions and acknowledgements no longer solely captures participants' attention pattern as it refers to understanding the state of the system, since it removes the distinction between information acquisition and control activities. Still, this could prove to be a useful method for characterizing more general attention patterns and future research is encouraged on this topic.

Limitations

The Rancor Microworld is intended to serve as a research tool for examining human factors issues and the related psychological constructs, such as SA and attention. The microworld embodies a very simple representation of the basic system layout, system dynamics, and general operating practices found at existing NPPs. As such, the microworld is most accurately described as a low-fidelity simulator with a significantly reduced scope and complexity. This reduced fidelity results in many systems and subcomponents of the plant not being included, and as a result, the generalizability is limited. The neutronics, pressurizer, chemical and volume control system, main steam reheaters, feedwater reheaters, turbine control system, and electrical distribution are not included in this implementation of the microworld. The control systems are also highly simplified and lack much of the automation typically found in existing NPPs. Furthermore, some of the physical processes are simplified, such as the speed control of the turbine. Including these additional systems and enhancing the fidelity would significantly increase the complexity and undermine the goals of maintaining sufficient simplicity that a novice student population could be quickly trained to use the simulation. Research examining low- and high-fidelity simulators in healthcare suggests the amount of fidelity should align with the expertise of the participants (Munshi, Lababidi, & Alyousef, 2015). As simulation fidelity increases, the complexity can overwhelm participants, while in contrast, as fidelity is reduced, the participant may become skeptical of the simulations' validity. The fidelity of the Rancor Microworld, though certainly on the lower end of the spectrum, strikes a balance necessary for students and the expert population sample comprised of steam plant operators. This reduced fidelity affords the use of novice student participants and greatly reduces the reliance on SMEs with plant-specific knowledge typically required for performing full-scope simulation experiments (Ulrich, Werner, & Boring, 2015). In regard to generalizability, any research must be explicit in reporting the results of low-fidelity simulation experiments to make it clear as to what aspects can and cannot be reliably generalized to an expert population and real-world situations. Furthermore, results from these low-fidelity experiments should be verified with full-scope simulations and actual operator participants to understand the similarities and differences.

The operators used in this study were not, in fact, licensed NPP control room operators. The steam plant operators used in this study underwent less training, and therefore, could be classified as less expert than actual NPP operators. The challenge of recruiting and working with actual NPP operators cannot be emphasized enough. NPP operators represent one of the most difficult populations to sample. Future research on NPP operators and other types of process control experts are needed to further understand differences between novice students and expert populations on process control tasks, such as the microworld used in these studies.

Future Directions

These initial research efforts have demonstrated the use of the Rancor Microworld as a research platform for examining psychological principles associated with SA and human performance in a simulated complex process control system. The overall task that participants experienced was centered on configuring the plant for an online mode of operation to produce power, and then recover from a scripted fault. Participants were instructed to generate as much power as possible to maximize their score. Though this was a useful approach for evaluating the microworld and providing a task with sufficient complexity to keep the participant engaged and induce, to varying extents, some variability in the various performance, attention, and SA measures, future studies should adopt a more explicit power-production goal.

A single type of acknowledgement marker was used to assess attention in this line of research. The marker selected for these studies was evaluated and vetted in a series of pilot studies to ensure that it did not overly attract participants' attention while they were attending to other locations, but once attention was allocated near its location, they could easily discriminate its state of motion and perform the acknowledgement. The primary goal of this research was to develop a practical method for assessing attention during a complicated simulation task and this marker demonstrated the requisite attentional properties to measure attention patterns. Furthermore, the results provide supporting evidence that the acknowledgement markers were successful in measuring attention similar to that of eye tracking fixations in over longer time-periods and situations that did not require a large number of control actions. However, the marker used for this line of research is certainly not the only marker stimulus that could be used as the basis for the acknowledgement marker task. Other marker stimuli should be explored to refine the measure and further mitigate unwanted attention capture effects while also retaining detectability. Additionally, the burden of clicking the marker to respond as an acknowledgement of the marker in the target, rotational motion state, could be reduced by adjusting the size of the hitbox. The hitbox refers to the area over and around the marker stimulus that registers mouse clicks. Larger hitboxes would reduce the effort required to accurately select the necessary region to register an acknowledgement in accordance with Fitts' Law (1954). One potential way to increase the hitbox entails exploring some drastically different stimuli as the marker. For example, the marker could be embedded within the border of indicators as a moving dashed line. The hitbox could be expanded to cover the entire control using this border stimulus and increase the likelihood for capturing more acknowledgements, since the effort required to do so would be reduced. Future research is encouraged to explore alternative acknowledgement marker stimulus presentations.

Chapter 8. References

Anderson, S. J., Mullen, K. T., and Hess, R. F. (1991). Human peripheral spatial resolution for achromatic and chromatic stimuli: Limits imposed by optical and retinal factors. *Journal of Physiology*, 442, 47–64.

Bell H. H., Lyon D. R. (2000). Using observer ratings to assess situation awareness. In: Endsley M. R., Garland D. J. (Eds.), *Situation Awareness Analysis and Measurement* (pp. 129–146). Mahwah, NJ, USA: Lawrence Erlbaum Associates.

Boring, R. L., Agarwal, V., Joe, J. C., & Persensky, J. J. (2012a). Digital Full-Scope Mockup of a Conventional Nuclear Power Plant Control Room, Phase 1: Installation of a Utility Simulator at the Idaho National Laboratory. INL/EXT-12-26367. Idaho Falls, ID, USA: Idaho National Laboratory.

Boring, R. L., Ulrich, T. A., Joe, J. C., & Lew, R. T. (2015). Guideline for operational nuclear usability and knowledge elicitation (GONUKE). *Procedia Manufacturing*, 3, 1327–1334.

Boring, R., Agarwal, V., Fitzgerald, K., Hugo, J., & Hallbert, B. (2013). Digital Full-Scope Simulation of a Conventional Nuclear Power Plant Control Room, Phase 2: Installation of a Reconfigurable Simulator to Support Nuclear Plant Sustainability. INL/EXT-13-28432. Idaho Falls, ID, USA: Idaho National Laboratory.

Boring, R., Kelly, D., Smidts, C., Mosleh, A., & Dyre, B. (2012b). Microworlds, simulators, and simulation: Framework for a benchmark of human reliability data sources. In: *11th International Probabilistic Safety Assessment and Management Conference and the*

Annual European Safety and Reliability Conference 2012 (PSAM11 ESREL 2012), 25–29 June, Helsinki, Finland. 16B-Tu5-5.

Boring, R., Lew, R., & Ulrich, T. (2017). Advanced Nuclear Interface Modeling Environment (ANIME): A tool for developing human-computer interfaces for experimental process control systems. In: Nah, F., & Tan, C. (Eds.). *Proceedings of the 4th International Conference on HCI in Business, Government, and Organizations (HCIBGO 2017)* 9–14 July. Vancouver, BC, Canada (pp. 3-15). *Lecture Notes in Computer Science*, 10293. Cham, Switzerland: Springer.

Broadbent, D. E. (1958). *Perception and Communication*. New York, NY, USA: Oxford University Press.

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

Cao, A., Chintamani, K. K., Pandya, A. K., & Ellis, R. D. (2009). NASA TLX: software for assessing subjective mental workload. *Behavior Research Methods*, 41(1), 113–117.

Carrasco, M., & McElree, B. (2001). Covert attention accelerates the rate of visual information processing. *Proceedings of the National Academy of Sciences*, 98(9), 5363–5367.

Carvalho, P. V., Vidal, M. C., & de Carvalho, E. F. (2007). Nuclear power plant communications in normative and actual practice: A field study of control room operators’ communications. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 17(1), 43–78.

Cave, K. R., & Bichot, N. P. (1999). Visuospatial attention: Beyond a spotlight model. *Psychonomic Bulletin & Review*, 6(2), 204–223.

Cave, K. R., & Kosslyn, S. M. (1989). Varieties of size-specific visual selection. *Journal of Experimental Psychology: General*, 118(2), 148–164.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, 25(5), 975–979.

Connor, C. E., Egeth, H. E., & Yantis, S. (2004). Visual attention: Bottom-up versus top-down. *Current Biology*, 14(19), R850–R852.

Demas, M., Lau, N., & Elks, C. (2015). Advancing human performance assessment capabilities for integrated system validation—A human-in-the-loop experiment. In: *Proceedings of the 9th American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation & Control and Human-Machine Interface Technologies (NPIC & HMIT)*. Charlotte, NC, USA.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining “gamification.” In: *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (MindTrek ’11)*. pp. 9-15. New York, NY, USA: ACM.

Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology*, 92(1), 53–78.

Duchowski, A. (2007). Eye tracking methodology: Theory and practice (Vol. 373). London, England: Springer Science & Business Media.

Durlach, P. J., Kring, J. P., & Bowens, L. D. (2008). Detection of icon appearance and disappearance on a digital situation awareness display. *Military Psychology*, 20(2), 81–94.

Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., Nikolic, D., & Manning, C. A. (1998). Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly*, 6(1), 1–20.

Dyre, B. P., Adamic, E. J., Werner, S., Lew, R., Gertman, D. I., & Boring, R. L. (2013). A microworld simulator for process control research and training. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 57(1) 1367–1371. Los Angeles, CA: SAGE Publications.

Endsley, M. R. (1995a). Towards a theory of situation awareness in dynamic systems. *Human Factors*, 37, 32–64.

Endsley, M. R. (1995b). Measurement of situation awareness in dynamic systems. *Human Factors*, 37, 65–84.

Endsley, M. R. (2000). Theoretical underpinnings of situation awareness: A critical review. In: Endsley M. R., Garland D. J. (Eds.), *Situation Awareness Analysis and Measurement* (pp. 3–32). Mahwah, NJ, USA: Lawrence Erlbaum Associates.

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. In: *Proceedings of the*

Human Factors and Ergonomics Society 42nd Annual Meeting. pp. 82–86. Santa Monica, CA, USA: SAGE Publications.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, 16(1), 143–149.

Eriksen, C. W., & Yeh, Y. Y. (1985). Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception and Performance*, 11(5), 583–597.

Fisk, A. D., & Schneider, W. (1981). Control and automatic processing during tasks requiring sustained attention: A new approach to vigilance. *Human Factors*, 23(6), 737–750.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6), 381–391.

Flach, J. M. (1995). Situation awareness: Proceed with caution. *Human Factors*, 37(1), 149–157.

Fletcher, C. D., & Schultz, R. R. (1995). RELAP5/MOD3 Code Manual, Volume V: User's Guidelines. INEEL/EXT-95-00837. Idaho Falls, ID, USA: Idaho National Laboratory.

Fracker, M. L. (1991). *Measures of situation awareness: Review and future directions* (No. AL-TR-1991-0128), Montoursville, PA, USA: Logue (George E) Inc.

Gaba, D. M., Howard, S. K., & Small, S. D. (1995). Situation awareness in anesthesiology. *Human Factors*, 37, 20–31.

Goldberg, J. H., & Wichansky, A. M. (2003). Eye tracking in usability evaluation: A practitioner's guide. In: Hyona, J., Radach, R., & Duebel, H. (Eds.), *The Mind's Eye:*

Cognitive and Applied Aspects of Eye Movement Research (pp. 573–605). Amsterdam: Elsevier Science.

Graneheim, U. H., & Lundman, B. (2004). Qualitative content analysis in nursing research: Concepts, procedures, and measures to achieve trustworthiness. *Nurse Education Today*, 24(2), 105–112.

Gugerty, L. J. (1997). Situation awareness during driving: explicit and implicit knowledge in dynamic spatial memory. *Journal of Experimental Psychology: Applied*, 3, 42–66.

Harms, L., & Bundesen, C. (1983). Color segregation and selective attention in a nonsearch task. *Attention, Perception, & Psychophysics*, 33(1), 11–19.

Hart, S. G. (2006). NASA-task load index (NASA-TLX): 20 years later. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9), 904–908. Los Angeles, CA, USA: Sage Publications.

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

Hartman, B. O., & Secrist, G. E. (1991). SA is more than exceptional vision. *Aviation, Space, and Environmental Medicine*, 62, 1084–1089.

Hauland, G. (2003). Measuring team situation awareness by means of eye movement data. In: *Proceedings of HCI International 2003: Vol 3* (pp. 230–234). Mahwah, NJ, USA: Lawrence Erlbaum Associates.

Hogg, D. N., Folles, 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.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). Eye tracking: A comprehensive guide to methods and measures. Oxford, UK: Oxford University Press.

Hoonakker, P., Carayon, P., Gurses, A. P., Brown, R., Khunlertkit, A., McGuire, K., & Walker, J. M. (2011). Measuring workload of ICU nurses with a questionnaire survey: The NASA Task Load Index (TLX). *IIE Transactions on Healthcare Systems Engineering*, 1(2), 131–143.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277–1288.

Hsu, Y. C. (2006). The effects of metaphors on novice and expert learners' performance and mental-model development. *Interacting with Computers*, 18(4), 770–792.

Ishai, A., & Sagi, D. (1995). Common mechanisms of visual imagery and perception. *Science*, 268(5218), 1772–1774.

Isreal, J. B., Chesney, G. L., Wickens, C. D., & Donchin, E. (1980). P300 and tracking difficulty: Evidence for multiple resources in dual-task performance. *Psychophysiology*, 17(3), 259–273.

Itti, L. (2000). Models of bottom-up and top-down visual attention. Doctoral dissertation, California Institute of Technology.

Itti, L., Braun, J., Lee, D. K., & Koch, C. (1998). A model of early visual processing. In: Jordan, M. I., Kearns, M. J., Solla, S. A. (Eds.), *Advances in Neural Information Processing Systems (NIPS*1997)*, (pp. 173–179). Cambridge, MA, USA: MIT Press.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11),

1254–1259.

James, W. (1890). *The Principles of Psychology*. Volumes I and II. New York, NY, USA: Holt.

Ji, Q., & Yang, X. (2002). Real-time eye, gaze, and face pose tracking for monitoring driver vigilance. *Real-Time Imaging*, 8(5), 357–377.

Jonides, J. (1981). Voluntary versus automatic control over the mind's eye's movement. In: Long, J. B., & Baddeley, A. D. (Eds.), *Attention and Performance IX* (pp. 187–203). Hillsdale, NJ, USA: Lawrence Erlbaum Associates.

Kaber, D. B., & Endsley, M. R. (1998). Team situation awareness for process control safety and performance. *Process Safety Progress*, 17(1), 43–48.

Kahneman, D. (1973). Attention and Effort. Englewood Cliffs, NJ, USA: Prentice-Hall.

Kasarskis, P., Stehwien, J., Hickox, J., Aretz, A., & Wickens, C. (2001). Comparison of expert and novice scan behaviors during VFR flight. In: *Proceedings of the 11th International Symposium on Aviation Psychology* (pp. 1–6). Columbus, Ohio, USA: Ohio State University.

Kelly, M. J., Wooldridge, L., Hennessey, R. T., Vreuls, D., Barnebey, S. F., Cotton, J. C., & Reed, J. C. (1979). Air Combat Maneuvering Performance Measurement, AFHRL-TR-79-3. September 1979. San Antonio, TX, USA: Air Force Human Resources Laboratory, Brooks Air Force Base.

Kemeny, J. G. (1979). *Report of the President's Commission on the Accident at Three Mile Island—The Need for Change: The Legacy of TMI*. Washington, DC, USA: U.S. Government Printing Office.

Kimball, D. R., & Holyoak, K. J. (2000). Transfer and expertise. In: Tulving, E., & Craik, F. I. M. (Eds.), *Oxford Handbook of Memory* (pp. 109–122) London, UK: Oxford University Press.

Kovesdi, C. R., Rice, B. C., Bower, G. R., Spielman, Z. A., Hill, R. A., & Le Blanc, K. L. (2015). Measuring Human Performance in Simulated Nuclear Power Plant Control Rooms Using Eye tracking. INL/EXT-15-37311. Idaho Falls, ID, USA: Idaho National Laboratory.

Kovesdi, C., Barton, B. K., & Rice, L. (2012). Visual efficiency-detection index: A new composite measure of visual search. *Journal of Eye tracking, Visual Cognition and Emotion*, 2. ISSN: 1647–7677.

Kowler, E. (2011). Eye movements: The past 25 years. *Vision Research*, 51(13), 1457–1483.

Kravitz, D. J., & Behrmann, M. (2011). Space-, object-, and feature-based attention interact to organize visual scenes. *Attention, Perception, & Psychophysics*, 73(8), 2434–2447.

LaBerge, D. (1983). Spatial extent of attention to letters and words. *Journal of Experimental Psychology: Human Perception and Performance*, 9(3), 371–379.

Larsen, A., & Bundesen, C. (1978). Size scaling in visual pattern recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 4(1), 1–20.

Law, B., Atkins, M. S., Kirkpatrick, A. E., & Lomax, A. J. (2004). Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment. In: *Proceedings of the 2004 Symposium on Eye tracking Research and Applications (ETRA '04)* (pp. 41–48). New York, NY, USA: ACM Press.

Lee, J. D., Caven, B., Haake, S., & Brown, T. L. (2001). Speech-based interaction with in-vehicle computers: The effect of speech-based e-mail on drivers' attention to the roadway. *Human Factors*, 43(4), 631–640.

Lew, R., Boring, R. L., & Ulrich, T. A. (2014). A prototyping environment for research on human-machine interfaces in process control use of Microsoft WPF for microworld and distributed control system development. In: *2014 7th International Symposium on Resilient Control Systems (ISRCS)*, pp. 1–6. IEEE, August 2014.

Mackworth, J. F. (1968). Vigilance, arousal, and habituation. *Psychological Review*, 75(4), 308–322.

Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, 1(1), 6–21.

Marquart, G., Cabrall, C., & de Winter, J. (2015). Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing*, 3, 2854–2861.

Matthews, M. D., & Beal, S. A. (2002). Assessing Situation Awareness in Field Training Exercises. U.S. Army Research Institute for the Behavioral Sciences. Research Report 1795.

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual skills in airport-security screening. *Psychological Science*, 15(5), 302–306.

McHugh, M. L. (2012). Inter-rater reliability: The kappa statistic. *Biochemia Medica*, 22(3), 276–282.

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, 11(1), 56–60.

Munshi, F., Lababidi, H., & Alyousef, S. (2015). Low- versus high-fidelity simulations in teaching and assessing clinical skills. *Journal of Talibah University Medical Sciences*, 10, 12–15.

Nevin, J. A. (1969). Signal detection theory and operant behavior: A review of David M. Green and John A. Swets' signal detection theory and psychophysics. 1. *Journal of the Experimental Analysis of Behavior*, 12(3), 475–480.

Nielsen, J. (1989). Usability engineering at a discount. In: Salvendy, G., & Smith, M. J. (Eds.), *Designing and Using Human-Computer Interfaces and Knowledge Based Systems*, (pp. 394–401). Amsterdam: Elsevier Science Publishers.

Nobre, A. C., & Kastner, S. (Eds.). (2014). *Oxford Handbook of Attention*. New York, NY, USA: Oxford University Press.

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

O’Hara, J., Higgins, J., Stubler, W., Goodman, C., Eckinrode, R., Bongarra, J., & Galletti, G. (2011). Human factors engineering program review model. NUREG-0711, rev. 3. Washington, DC, USA: U. S. Nuclear Regulatory Commission.

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, 116(2), 220–244.

Patrick, J., James, N., Ahmed, A., & Halliday, P. (2006). Observational assessment of situation awareness, team differences, and training implications. *Ergonomics*, 49(4), 393–417.

Poole, A., Ball, L. J., & Phillips, P. (2005). In search of salience: A response-time and eye-movement analysis of bookmark recognition. In Fincher, S., Markopoulos, P., Moore, D., & Ruddle, R. (Eds.), *People and Computers XVIII—Design for Life* (pp. 363–378). London, UK: Springer.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32(1), 3–25.

Posner, M. I., Snyder, C. R. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, 109, 160–174.

Prinzmetal, W., McCool, C., & Park, S. (2005). Attention: Rreaction time and accuracy reveal different mechanisms. *Journal of Experimental Psychology: General*, 134(1), 73–92.

Ragsdale, A., Lew, R., & Boring, R. (2015). A study on trust in alarms in a nuclear power plant microworld simulation. *2015 Resilience Week*, 18–20 August, Philadelphia, PA, USA.

Rizzolatti, G., Riggio, L., Dascola, I. and Umiltá C. (1987). Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25, 31–40.

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

Salmon, P., Stanton, N., Walker, G., & Green, D. (2006). Situation awareness measurement: A review of applicability for C4i environments. *Applied Ergonomics*, 37(2), 225–238.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye tracking protocols. In: *Proceedings of the 2000 Symposium on Eye tracking Research & Applications (ETRA '00)* 6–8 November. Palm Beach Gardens, FL, USA (pp. 71–78). New York, NY, USA: ACM.

Sarter, N. B., & Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology*, 1(1), 45–57.

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*, 37(1), 5–19.

Sarter, N. B., Mumaw, R. J., & Wickens, C. D. (2007). Pilots' monitoring strategies and performance on automated flight decks: An empirical study combining behavioral and eye tracking data. *Human Factors*, 49(3), 347–357.

Senge, P. M. (2006). *The Fifth Discipline: The Art and Practice of the Learning Organization*. New York, NY, USA: Doubleday.

Shaw, M. L., & Shaw, P. (1977). Optimal allocation of cognitive resources to spatial location. *Journal of Experimental Psychology: Human Perception & Performance*, 3(2), 201–211.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, 28(9), 1059–1074.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, 1(7), 261–267.

Smith, K., & Hancock, P. A. (1995). Situation awareness is adaptive, externally directed consciousness. *Human Factors*, 37(1), 137–148.

Strayer, D. L., & Johnston, W. A. (2001). Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological Science*, 12(6), 462–466.

Strayer, D. L., Drews, F. A., & Johnston, W. A. (2003). Cell phone-induced failures of visual attention during simulated driving. *Journal of Experimental Psychology: Applied*, 9(1), 23–32.

Taylor, R. M. (1990). Situation Awareness Rating Technique (SART): The development of a tool for aircrew systems design (AGARD-CP-478). In: *Situation Awareness in Aerospace Operations* NATO-AGARD, Neurilly Sur Seine, France (SEE N 90-28972 23-53)

(pp. 3/1–3/17).

Telford, C. (1931). The refractory phase of voluntary and associative response. *Journal of Experimental Psychology*, 14, 1–35.

Tobii Technology AB. (2016). *Accuracy and precision test report X2-60 fw 1.0.5*. Sweden, Karlsrovägen: Tobii Techology AB, LLC.

Treisman, A. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, 12, 242–248.

Treisman, A. (1969). Strategies and models of selective attention. *Psychological Review*, 76, 282–299.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97–136.

Ulrich, T. A., Lew, R., Boring, R. L., & Thomas, K. (2014). A computerized operator support system prototype. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1), 1899–1903. Los Angeles, CA, USA: SAGE Publications.

Ulrich, T. A., Lew, R., Werner, S., & Boring, R. (in press). Rancor: A gamified microworld nuclear power plant simulation for engineering psychology research and process control applications. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Los Angeles, CA, USA: SAGE Publications.

Ulrich, T. A., Werner, S., & Boring, R. L. (2015). Studying situation awareness on a shoestring budget: An example of an inexpensive simulation environment for theoretical research. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59(1), 1520–1524. Los Angeles, CA, USA: SAGE Publications.

Ulrich, T., Boring, R., & Lew, R. (2014). Human Factors Engineering Design Phase Report for Control Room Modernization. INL/EXT-14-33221. Idaho Falls, ID, USA: Idaho National Laboratory.

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

Ulrich, T., Werner, S., Lew, R., & Boring, R. (2016). COSSplay: Validating a computerized operator support system using a microworld simulator. In: *International Conference on Human-Computer Interaction* (pp. 161–166). Springer International Publishing Switzerland.

United States Nuclear Regulatory Commission (1998, April). NRC staff Proposes $55,000 Fine Against Consumers Power Company for October Incident at Palisades Nuclear Plant. https://www.nrc.gov/docs/ml0037/ML003707614.pdf.

United States Nuclear Regulatory Commission (2017, February). List of Power Reactor Units. https://www.nrc.gov/reactors/operating/list-power-reactor-units.html.

United States Nuclear Regulatory Commission. (2001). Nuclear Power Plant Simulation Facilities for Use in Operator Training and License Examinations, Rev. 3, Regulatory Guide 1.149. Washington, DC, USA: U.S. Nuclear Regulatory Commission.

Van der Heijden, A. H. C. (1996). Perception for selection, selection for action, and action for perception. *Visual Cognition*, 3(4), 357–361.

Vicente, K. J., Christoffersen, K., & Pereklita, A. (1995). Supporting operator problem solving through ecological interface design. *IEEE Transactions on Systems, Man, and Cybernetics*, 25(4), 529–545.

Vicente, K. J., Moray, N., Lee, J. D., Hurecon, J. R., Jones, B. G., Brock, R., & Djemil, T. (1996). Evaluation of a Rankine cycle display for nuclear power plant monitoring and diagnosis. *Human Factors*, 38(3), 506–521.

Vicente, K., & Pawlak, W. (1994). Cognitive work analysis for the DURESS II system. *Cognitive Engineering Laboratory, Department of Industrial Engineering, Toronto, Canada CEL*, 94–93. University of Toronto.

Waag, W. L., & Houck, M. R. (1994). Tools for assessing situation awareness in an operational fighter environment. Aviation, space, and environmental medicine.

Waern, Y., & Cañas, J. (2003). Microworld task environments for conducting research on command and control. *Cognition, Technology, & Work*, 5(3), 181–182.

Walker, G. H., Stanton, N. A., & Young, M. S. (2008). Feedback and driver situation awareness (SA): A comparison of SA measures and contexts. *Transportation Research Part F: Traffic Psychology and Behaviour*, 11(4), 282–299.

Wickens, C. D. (1991). Processing resources and attention. In: Damos, D. L. (Eds.), *Multiple-task Performance*, (pp. 3–34). Bristol, PA, USA: Taylor & Francis.

Wickens, C. D., & McCarley, J. S. (2008). *Applied Attention Theory*. Boca Raton, FL, USA: CRC Press.

Wickens, C. D., Alexander, A. L., Horrey, W. J., Nunes, A., & Hardy, T. J. (2004). Traffic and flight guidance depiction on a synthetic vision system display: The effects of clutter on performance and visual attention allocation. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 48(1) 218–222. Los Angeles, CA: SAGE Publications.

Wickens, C. D., Goh, J., Helleberg, J., Horrey, W. J., & Talleur, D. A. (2003). Attentional models of multitask pilot performance using advanced display technology. *Human Factors*, 45(3), 360–380.

Wickens, C. D., McCarley, J. S., Alexander, A. L., Thomas, L. C., Ambinder, M., & Zheng, S. (2008). Attention-situation awareness (A-SA) model of pilot error. *Human Performance Modeling in Aviation*, 213–239.

Wiley, J. (1998). Expertise as mental set: The effects of domain knowledge in creative problem solving. *Memory & Cognition*, 26(4), 716-730.

Wolfe, J. M. (1994). Visual search in continuous, naturalistic stimuli. *Vision Research*, 34(9), 1187–1195.