U.S.NRC
United States Nuclear Regulatory Commission
Protecting People and the Environment

RIL 2022-11

HUMAN PERFORMANCE TEST FACILITY (HPTF)

VOLUME 1 - SYSTEMATIC HUMAN PERFORMANCE DATA COLLECTION USING NUCLEAR POWER PLANT SIMULATOR: A METHODOLOGY

Manuscript Completed:

Date Published: January, 2023

Prepared by:

N. Hughes, A. D'Agostino, & K. Dickerson

G. Matthews*, L. Reinerman-Jones*, D. Barber*, J. Mercado*, J. Harris*, J. Lin*

*University of Central Florida

Niav Hughes Green, NRC Project Manager

Research Information Letter

Office of Nuclear Regulatory Research

Disclaimer

Legally binding regulatory requirements are stated only in laws, NRC regulations, licenses, including technical specifications, or orders; not in Research Information Letters (RILs). A RIL is not regulatory guidance, although NRC's regulatory offices may consider the information in a RIL to determine whether any regulatory actions are warranted.

PREFACE

HPTF RIL Series (RIL 2022-11) Preface

Much of the basis for current NRC Human Factors Engineering (HFE) guidance comes from data from research conducted in other domains (e.g., aviation, defense), qualitative data from operational experience in NPPs, and a limited amount from empirical studies in a nuclear environment. The Commission, in SRM SECY-08-0195, approved the staff's recommendation and directed the staff to consider using generic simulator platforms for addressing human performance issues, as simulators provide a tool to gather more empirical nuclear specific human performance data. These data would enhance the current information gathering process, thus providing stronger technical bases and guidance to support regulatory decision making. the former Office of New Reactors (NRO) issued a user need for the Office of Nuclear Regulatory Research (RES) to update its human factors (HF) review guidance with regards to emerging technologies (User Need NRO-2012-007) and more recently the Office of Nuclear Reactor Regulation (NRR) issued a follow-on user need with the same purpose (User Need NRR-2019-008). In the spring of 2012, the NRC sponsored a project to procure a low-cost simulator to empirically measure and study human performance aspects of control room operations to address the human performance concerns related to current as well as new and advanced control room designs and operations. Using this simulator, the Human Factors and Reliability Branch (HFRB) in the RES Division of Risk Assessment (DRA) began a program of research known as the NRC Human Performance Test Facility (HPTF) to collect empirical human performance data with the purpose of measuring and ultimately better understanding the various cognitive and physical elements that support safe control room operation. Additionally, the baseline methodology documented in these volumes will enable HRA data research that will address key gaps in available data for topics such as dependency and errors of commission, improving the state of the art of human reliability analysis (HRA) and thus dual HF and HRA data missions.

Recognizing the essential role of data to our HF and HRA programs, the NRC historically approached data collection through multiple avenues – all with their inherent strengths and weaknesses:

1. Licensed Operators – controlled experiments at the Halden Reactor Project
2. Licensed Operators – the Scenario Authoring, Characterization, and Debriefing Application (SACADA) database capturing training scenarios
3. Novice populations – scientific literature, laboratory settings – non-nuclear

The HPTF program captures data from both novice and operational populations and the work is specifically targeted to the nuclear domain. In addition, the HPTF methodology expands upon these data collection methods. Most notably, though the addition of a new population category, that of formerly licensed operators and other nuclear domain experts. The HPTF methodology (described in detail in RIL 2022-11 Volume 1) enables the NRC to fill in the gaps from the other 3 data collection activities and conduct responsive research to support the informational needs of our users (e.g., NRR HFE technical reviewers and HRA analysts).

The intent of the HPTF was to design experiments that balanced domain realism and laboratory control sufficiently to collect systematic, meaningful, human performance data related to execution of common nuclear main control room (MCR) tasks. Three large-scale experiments were conducted to address challenges associated with developing a research methodology for using novices in a highly complex, expert driven domain. These three experiments are reported

as Studies 1 and 2 in RIL 2022-11 Volume 1 which describes the approach and methodology underlying this research effort and the resulting findings for the series of studies. In RIL 2022-11 Volume 2, the Volume 1 findings were further validated via a fourth data collection by testing a formerly licensed operator population using a full-scale, full-scope simulator. Cross-experiment comparisons were enabled by leveraging a formerly licensed operator as a member of the research team to serve as senior reactor operator (SRO) and ensure participants received an experience as similar and structured as possible to the studies in Volume 1[1].

To ensure the developed methodology continues to support the HFE technical staff in user offices, the HPTF team works with those stakeholders to establish research questions and experimental design options for follow-on work. The experimental design and research questions that were examined were determined through a collaborative effort between NRC staff and a contractor with an identical simulator and performance assessment capabilities.

Toward this end, to date, three experimental design workshops have been held. The first workshop was held on March 5 and 6, 2018 upon completion of the first three HPTF experiments. The direction resulting from this first workshop was to validate the methodology and generalize the findings from the baseline HPTF experiments by using formerly licensed operators as participants to complete an experimental scenario using an analog, full-scope, full-scale simulator and a digital, part-task simulator. RIL 2022-11 Volume 2 describes the research approach and findings for the fourth experiment in the series.

The second workshop was held on August 20 and 21, 2019. The direction resulting from this second workshop was to perform a reanalysis of all HPTF experiments thus far to investigate: 1) Workload Measure Sensitivities 2) Task Order Effects and 3) Touchscreen Ergonomics. The results of each of these supplementary analyses and their regulatory implications are discussed in RIL-2022-11 Volumes 3-5 (in press). Due to the COVID-19 health crisis, the third workshop was held as a virtual series consisting of six 2-hour blocks between October 29 to November 20, 2020. The future direction topics discussed during the most recent workshop are described in RIL 2022-11 Volume 6 (in press). The final direction and experimental design are yet to be set, but the resulting methodology and results may be published as Volume 7.

These volumes of research illustrate the NRC's ongoing effort to perform systematic human performance data collection using a simulator to better inform NRC guidance and technical bases in response to Staff Requirements Memorandum (SRM) SECY-08-0195 and SRM-M061020. The HF and HRA data are essential to ensure that our HFE guidance documents and HRA methods support the review and evaluation of "state-of-the-art" HF programs (as required by 10 Code of Federal Regulations (CFR) 50.34(f)(2)(iii)).

[1] Systematic experimentation is challenging in the nuclear domain using real operators and full, dynamic scenarios because operators can take many paths to achieving a successful outcome. This variability represents a condition that is not conducive to controlled laboratory study. By including a confederate SRO in the study using a dynamic scenario, this hard to control variability is managed, thereby, enabling stable observations. See RIL 2022-11 Volumes 1 and 2 for examples of these methodological benefits.

ABSTRACT

The human factors engineering (HFE) staff of the U.S. Nuclear Regulatory Commission (NRC) is responsible for reviewing the safety of control room designs per 10 CFR 50.34(f)(2)(iii). Due to rapid technological improvements, advanced control room designs and upgrades to existing control rooms must be reviewed. Much of the HFE guidance supporting these reviews is based on surrogate domains (e.g., aviation, military) which may not generalize to the highly complex nuclear domain. To address concerns over limited generalizability, the Commission, in a Staff Requirements Memorandum (SRM) SECY-08-0195, directed the use of generic simulator platforms for examination of human performance issues. In response to the SRM and to support identification and analysis of potential human factors issues with modern technologies, the NRC sponsored a project to procure low-cost simulators to systematically measure and study the cognitive, and physical aspects of human performance during control room operations.

Even with a low-cost simulator, gathering data from operating crews can be expensive and difficult. To meet this challenge, the human performance test facility (HPTF) partnered with the University of Central Florida (UCF) to create new methods for training and testing novice participants (university students) to play the role of reactor operators (RO) in a simplified, yet realistic, setting.

This report documents two large-scale experiments using a generic simulator to assess the impact of different tasks and display types on subjective and physiological measures of workload. The use of novice participants required a focus on rule- and skill-based tasks instead of knowledge-based tasks, since the novices would not possess the required domain knowledge. The three types of tasks examined were: Checking, Detection, and Response Implementation (see page 2-6 for task definitions).

The UCF simulator was a digital representation of a generic analog nuclear power plant (NPP) Main Control Room (MCR) interface. Across all the experimental sessions, regardless of interface type or operator role, the detection task was always the most difficult and produced the highest workload of the three tasks. Subtle variations in the workload and participant performance as a function of display type (desktop vs. touchscreen) and task (detection, checking, and response implementation) are discussed in detail in the sections that follow. Overall, the results of these studies demonstrate that display type changes tend to drive differences in measures of workload associated with spatial processing, and that workload and performance measures can diverge depending on task parameters. These studies also demonstrate the feasibility of the HPTF at UCF and the value of using novice, readily available participants for evaluating performance on rule- and skill-based aspects of RO tasks.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The staff of the U.S. Nuclear Regulatory Commission (NRC) is responsible for reviewing and determining the acceptability of new reactor designs to ensure they support safe plant operations (10 CFR 50.34 (f)(2)(iii)). Human performance is a key component in the safe operation of Nuclear Power Plants (NPPs) (NRC, 2002). The human operator is a vital part of plant safety; thus, the NRC staff must understand the potential impact of new designs on human performance to make sound regulatory decisions. Much of the basis for current NRC Human Factors Engineering (HFE) guidance comes from research conducted in other domains (e.g., aviation, defense), qualitative data from operational experience in NPPs, and a limited number of empirical studies in a nuclear environment. For new designs, technologies, and concepts of operations, there is even less information. To address this information gap, the Commission in a Staff Requirements Memorandum (SRM) SECY-08-0195 directed the staff to consider using generic simulator platforms for addressing human performance issues. A simulator provides a means to gather empirical nuclear-specific human performance data that is targeted to enhancing the current information gathering process and providing stronger technical bases and guidance to support regulatory decision making.

The simulator used to address the information gap digitally represents analog instrumentation and controls (I&C) for a generic Westinghouse 3-Loop Pressurized Water Reactor controls (developed by GSE Power Systems). Using this simulator, the Human Factors and Reliability Branch (HFRB) in the Office of Nuclear Regulatory Research (RES) launched a program of experimental research with the help of the Human Performance Test Facility (HPTF) to collect empirical human performance data for measuring and understanding the various cognitive and physical elements that support safe control room operation. The intent was to design experiments that balanced domain realism and laboratory control sufficiently to collect systematic meaningful human performance data related to execution of common main control room (MCR) tasks. Investigators identified and defined three types of tasks central to the MCR: Checking, Detection, and Response Implementation. A variety of subjective and physiological measures were collected to understand the performance of those tasks in terms of both physiological and subjective workload.

Chapter 1 of this report introduces the research topic and provides background motivation and project history including the challenges faced by the research team and the methodological choices that influenced the project direction. Chapter 2 gives an overview of the research approach including the research questions which framed the overall direction and a description of the proof of concept and the general methods that were used for the series of experiments. Chapter 3 reports the results of Study 1 which 1) established the methodology or proof of concept (i.e., simplified yet similar environment and tasks, training novice participants), 2) examined novice workload level and type associated with task types, and 3) examined differences in workload and task performance associated with two different interfaces. Chapter 4 reports the results of Study 2 which used the same method (i.e., same simulated environment, same task types and a multivariate assessment of workload) with formerly licensed nuclear power plant (NPP) or nuclear submarine operators as participants to further validate the methodology and confirm the generalizability of workload response trends to the nuclear domain. Chapter 5 includes a general discussion and conclusions for the research program thus far and proposes future directions for the program of research based on these conclusions. Future work using the HPTF paradigm will enable the fulfillment of the spirit and eventually the letter of the Commission's SRM which served as the catalyst for this research.

ABBREVIATIONS AND ACRONYMS

ANOVA	Analysis of Variance
BWR	Boiling Water Reactor
CBFV	Cerebral Blood Flow Velocity
CFR	Code of Federal Regulation
ECG	Electrocardiogram
EEG	Electroencephalogram
EOP	Emergency Operating Procedure
fNIR	Functional Near-Infrared Spectroscopy
GPWR	Generic Pressurized Water Reactor
HFE	Human Factors Engineering
HFRB	Human Factors and Reliability Branch
HPTF	Human Performance Test Facility
HR	Heart Rate
HRA	Human Reliability Analysis
HRV	Heart Rate Variability
HSI	Human System Interface
IBI	Inter-beat Interval
I&C	Instrumentation and Control
ISA	Instantaneous Self Assessment
ISO	International Standards Organization
M	Mean
MCR	Main Control Room
Mdn	Median
MRQ	Multiple Resource Questionnaire
NPP	Nuclear Power Plant
NRC	U.S. Nuclear Regulatory Commission
OP	Operating Procedure
PSF	Performance Shaping Factor
PWR	Pressurized Water Reactor
RES	Office of Nuclear Regulatory Research
RO	Reactor Operator
SME	Subject Matter Expert
SRO	Senior Reactor Operator
TCD	Transcranial Doppler
TLX	Task Load Index
TTC	Technical Training Center

1 INTRODUCTION

The U.S. Nuclear Regulatory Commission (NRC) regulates 93 commercial operating reactors across the United States that produce nearly 19% of the country's electrical power (NRC, 2021). The staff of the NRC are responsible for reviewing and determining the acceptability of new reactor designs to ensure they support safe plant operations (10 CFR 50.34 (f)(2)(iii). Human performance is a key aspect of the safe operations of Nuclear Power Plants (NPPs) (NRC, 2002); thus, the NRC staff must understand the potential impact of new designs on human performance.

1.1 Background

The nuclear power industry is unique in terms of both age and complexity when compared to other industrial domains. For example, much of the operating fleet of reactors in the United States were built in the 1970s and 1980s and licensed to operate for 40 years. As we are beyond the end of this first cycle of operating licenses, most plants have applied to the NRC for 20-year license extensions and in some cases even applied for second 20-year license extensions. Due to obsolescence issues with older analog technologies, there is a growing need to assess the feasibility of integrating new digital technologies and understanding how digital modernization influences human reliability under routine and off-normal plant conditions. Further adding to the complexity of maintaining safe operations in an aging fleet is that modernization efforts often need to occur incrementally in coordination with outage schedules resulting in a trend towards hybrid control rooms (Hugo, Slay, Hernandez, 2017). Hugo et al., (2017, see also Lew et al., 2017) indicate that the result of this incremental modernization is digital I&Cs are replacing analog I&Cs for non-safety monitoring and control systems (e.g., feedwater or turbine control systems) while plants are maintaining analog I&Cs for safety systems.

While the legacy fleet is moving towards hybrid I&Cs, a number of plants with fully digital MCRs are scheduled to start-up within the next few years (NRC, 2020). These new digital plants have the potential to leverage tools like touchscreens and automation, which could confer benefits over conventional analog controlled plants. However, digital controls may pose new human factors challenges that will need to be considered by industry, designers, and regulators alike. Thus, they are being considered as an area of interest for systematic HFE investigation.

1.2 Program History and Development

The regulation most commonly associated with HFE is 10 CFR 50.34(f)(2)(iii)[2] which states:

> "Provide, for Commission review, a control room design that reflects state-of-the-art human factor principles prior to committing to fabrication or revision of fabricated control room panels and layouts."

The NRC's human factors technical staff use Chapter 18 of the Standard Review Plan (SRP) (NRC, 2016) to ensure that the regulations are met when performing a safety evaluation for

[2]For new reactor designs applying under Part 52, 10 CFR 52.47(a)(8) indicates that part 52 applicants must comply with certain parts of 10 CFR 50.34 including subpart (f).

license applications[3] for operating and new reactors (see 10 CFR Parts 50 and 52 respectively). The SRP references NUREG-0700, "Human-System Interface Design Review Guidelines" (NRC, 2002) and NUREG-0711,"Human Factors Engineering Program Review Model" (NRC, 2012), as guidance to support staff when applying the SRP (NRC, 2017).

NUREG-0711 (NRC, 2012) identified 12 review elements important to effective HFE in NPPs. Several of these elements identified the constructs of workload, situation awareness, and teamwork as important considerations during the design process. As such, applicants proposed a variety of metrics to measure these influences on human performance, with the goal being a demonstration of successful operator performance during testing at the design phase, prior to implementation. NUREG-0711 also requires an assessment of workload during Integrated System Validation (ISV) as part of a Validation and Verification (V&V) process. In the past, industry frequently relied on precedent when choosing the workload methods and metrics used to demonstrate the adequacy of their HFE programs. Precedent led to the overwhelming use of the subjective workload scale, the NASA Task Load Index (NASA-TLX: Hart & Staveland, 1988; Hart, 2006). Despite its ubiquity, the NASA-TLX may fail to capture elements of workload that could be identified using other methods, for example alternative subjective scales or objective methods based on recording physiological responses or performance. Given the complexity of the MCR environment, particularly a hybrid or digital MCR, it is likely that a multidimensional assessments of workload would be needed to fully capture the range of vulnerabilities that may threaten operator performance (see Matthews & Reinerman-Jones, 2017). Similar considerations likely apply to measurement of situation awareness and teamwork.

The challenge in implementing a multidimensional assessment strategy is that many of the alternatives and/or additions to the NASA-TLX were developed for use in other domains (e.g., military, aerospace, aviation) with specific populations (e.g., pilots, air traffic controllers). Theories of workload, situational awareness, and performance-based measures have the same potentially limited domain and population generalizability. The timescale of events and decisions and actions about those events is very different for pilots (milliseconds to seconds) compared to nuclear control room operators (hours to days). Based on currently available theories of workload and research on the temporal dimension of the NASA-TLX, it is not clear how different timescales influence previous reports of temporally-associated workload effects.

Other metrics, such as those captured by wired physiological equipment may be poorly suited or simply impractical as their can interfere with ongoing operations by restricting operator mobility. Consequently, the NRC published NUREG/CR-7190 which reviewed these metrics to determine the domains for which each metric was validated and each metric's strengths and limitations to supplement NRC technical reviewers' knowledgebase on the use of these metrics and to aid their evaluation of their proposed use in licensee applications (NRC, 2015).

The HPTF research program was established on the heels of NUREG/CR-7190 (NRC, 2015) and aims to provide the capability for empirical and systematic data collection to validate the constructs and measures of workload and situational awareness, as well as future concepts of operations, human reliability, performance, and other topics that may require novel data collection. The overall program is broad and aims to be responsive to agency needs. The first set of studies focuses on documenting the utility of and providing further validation for

[3]License applications can include construction permits, operating licenses, standard design certifications, and combined licenses

subjective, performance-based, and physiological measures of workload for use in the nuclear domain.

1.2.1 Rationale for Research Using a Low-Cost Simulator

In SECY-08-0195 the Commission directed the staff to consider using generic simulator platforms for examining human performance challenges. A simulator has the potential to enable collection of data specific to human performance in a nuclear-specific context. However, full scale simulators are costly and in limited supply, as are expert reactor operators (ROs). Historically, human factors research in the nuclear domain required purchasing an NPP MCR simulator and having a facility where all the "hard" analog panels can be staged. This kind of simulator also requires support; trained operations and IT staff are needed to use and maintain the simulator. This level of support requires a large start-up budget to build/buy the hardware, and a large operational budget for staffing and maintenance. This approach was not a financially viable option for the Office of Nuclear Regulatory Research (RES), as a result, the staff pursued several alternatives including:

- Collecting human performance data in the simulators at the NRC Technical Training Center (TTC)
- Partnering with a utility to collect data in their simulator
- Exploring availability of "soft" simulators (i.e., runs on a computer, no "hard" panels)

Collecting human performance data at the TTC and partnering with a utility were quickly ruled out because:

- Gaining access to either the TTC simulators or a utility simulator for research purposes including long periods of time for data collection is very difficult as they are often in use for training purposes.
- To operate a full-scope, full-scale simulator, trained NPP engineers and operators must be used. This is a problem, as mentioned previously, because the number of trained operators is limited, hence, their ability to be available for research is very restricted.
- Small sample size due to practical constraints and the limited opportunities for strict experimental controls would lead to an expensive and difficult to acquire dataset with limited quantitative value.

Exploring the availability of a "soft" simulator seemed like a viable path forward. However, the "soft" simulator needed to have several specific attributes to meet the requirements for empirically sound, but ecologically valid research. Simulators for training and those used for systematic laboratory research differ in their capabilities due to the unique needs of each use case. Table 1-1 outlines the specific attributes required to conduct empirically sound and ecologically valid research and maps those attributes onto the base feature set of the GSE generic pressurized water reactor (GPWR) simulator.

Table 1-1 Comparison between HPTF Requirements and GSE GPWR Features

Category	HPTF Requirements	GSE GPWR Base Feature Set
	Must be a generic (pre-built) model	
Hardware	Fidelity of the simulator must be high enough not to mislead an experienced operator into error in actions	Generic 3-loop Westinghouse PWR. Hard panel mimics used for HSI. System update time is at least 2 times/sec.
Software	Must model primary and secondary systems	

	Must include basic process models of reactor physics, thermohydraulic, and control systems	Real-Time Advanced Core and Thermohydraulic (RETACT) thermal hydraulics code
	Must allow for full-range of power operations	Capability to run full range of power operations
	HSI must either simulate current hard-wired control room bench boards or advanced control room workstations	Each operator station can access all control room soft panels
Experimenter Usability	Must have a straightforward method to configure the simulator to run in several modes (e.g., fully-simulated mode or a semi-manual mode)	Operator stations can be preconfigured to display specific panel sections
	Must allow the NRC to conduct real-time, human-in-the-loop simulations so that operator responses can be observed and assessed during scenarios of various initial conditions, plant behaviors, malfunctions, and transients	Over twenty initial conditions (can accommodate up to 200), The simulator is pre-loaded with 100s of malfunctions, Includes operating procedures for full range of operations, plant operating "curve book," and technical specifications, Allows for instrumentation failure
	Must have graphic tools to modify interfaces, as well as the ability to build additional graphic displays to study the impacts of new interface features or modifications on human performance	The graphics development tool allows for drag and drop user interface
	Interface configuration must be flexible so that the simulator allows one individual or a team of personnel to perform tasks	
	Must include an instructor station capable of simulation control, monitoring, and data visualization activities	The software includes a graphics tool, an instructor station, and a real-time executive program
Study Usability	Must provide ways to allow for non-operator participants to perform simplified control room operator tasks	
	Must operate on desktop computers under a Microsoft Windows environment	Runs on eight 24-inch LCD screens, 4 Dell Precision Workstations with Single Quad CPU
	Must have a data-logging system to collect human performance data and real-time plant parameter process values and exporting data to files in a format readable by Microsoft Excel	Contains real-time trending for data capture and logging, Data logs can be exported to Excel

In the spring of 2012, the Human Factors and Reliability Branch (HFRB) in RES procured two copies of the GSE GPWR 3-loop Westinghouse pressurized water reactors[4] Two copies were purchased with the intent to house one at NRC headquarters and use it primarily for scenario

[4]After an open, competitive bidding process and assessment of a variety of simulator options, the simulator that best fit the needs of the NRC was the GSE Power Systems Generic Pressurized Water Reactor (GSE GPWR).

testing and development. The other would be housed with a vendor that had access to a pool of novice participants.

The program is currently known as the NRC Human Performance Test Facility (HPTF). The HPTF aims to collect empirical human performance data with the purpose of gaining better understanding of the various cognitive and physical elements of performance to help guide decision making about future control room designs while also managing cost- and access-challenges.

1.2.2 Approach for Overcoming Cost and Participant Challenges

As mentioned previously, conducting human performance studies with current or former licensed operators is often time and cost prohibitive. To address these challenges, HPTF human performance studies were conducted in two steps.

- Step 1: Test non-operators with various combinations of scenarios, system conditions, and new technologies to identify leading human performance indicators. The results would allow researchers to identify safety-critical or error-prone contexts as well as identify measurement tools (i.e., measures of human performance) most sensitive to changes within this environmental context.
- Step 2: Using the insights from the first step, replicate the findings in full scale simulators and/or with ROs to test specific error-prone scenarios and further elucidate potential human factors issues related to new NPP control room designs.

1.2.3 Accessing Participants

Step 1 of the approach used in the HPTF is critical because accessing licensed NPP operators presents a major challenge to the conduct of successful simulator studies. Drawing conclusions from experimental data often requires large sample sizes, which is difficult to attain considering the limited availability of licensed operators. To overcome this barrier, the NRC technical staff determined that it would be necessary to source research participants from the general population (the validity and success of this recruitment approach is discussed in section 2.2, 3.3.3 and 5.1.2 respectively). To support this, the NRC issued a request for proposal (RFP) for a commercial contract to partner with an organization that could access or recruit large numbers of research participants and assist with experimental design. After an open competitive bidding process, the NRC partnered with the University of Central Florida's Institute for Simulation and Training (UCF IST)[5]. The second of the two GSE GPRW simulators is currently housed at UCF.

1.2.3.1 Challenges of Using a Novice Population

As is the case with many HFE research activities, the staff had to weigh the trade-offs between research design decisions, ecological validity, and generalizability. Access to a larger pool of research participants was critical for the project's success; however, this access came with some specific limitations. To collect data that would be generalizable to ROs, using novice operators, the research team determined that the environment needed to induce participants to experience the task complexity and cognitive demands experienced by trained operator, but without requiring all the knowledge and skills of a trained operator (Lackey et al., 2014;

[5]The first two 5-year Task Order agreements contracted UCF IST. The follow-on work is currently (in 2022) undergoing an open competitive bidding process.

Reinerman-Jones et al., 2013). In other words, the methodological approach should adhere to the principle of *different but equal*; the environment (e.g., interface, task) is different, but the cognitive demands and associated workload are the same (Hughes, D'Agostino, & Reinerman-Jones, 2017).

This report describes the first two studies in a series that will systematically collect human performance data for critical tasks in NPP MCRs through the design and execution of human-in-the-loop experiments. The research was conducted using a new experimental simulator based on the GSE Generic Pressurized Water Reactor (GPWR) platform and included a full complement of workload measures to produce a comprehensive assessment of workload for the nuclear domain. More specifically, a selection of a variety of subjective, objective (e.g., physiological), and performance-based measures of workload were included.

The objectives of the initial studies were to:

- Validate the methods and measures used during experimentation with the aim of establishing reasonable confidence that the results can be used to develop and inform the technical bases for NRC's HFE guidance.
- Establish baseline methods and a foundational data set for future studies.
- Determine some best practices for measuring task performance and operator workload.
- Profile and compare the workload types and levels associated with common tasks in a NPP MCR context using a multivariate assessment strategy that includes physiological, objective, and subjective measures of workload.

2 RESEARCH APPROACH AND GENERAL METHODS

The neural, sensory, and perceptual mechanisms underlying human cognition are complex and varied. These human factors impact performance influencing factors (NRC, 2012; 2016). While there is a challenge of generalizability of research from outside the nuclear domain, there is potential in generalizing research based on a general population to the reactor operator community. The underlying mechanisms of cognition are the same in operators as they are in a general population, therefore, it should be possible to use a population of novices as a proxy for expert operator (Hughes & D'Agostino, 2016). While it should be possible to use novices as proxies, the critical question is: *Can NPP MCR operations be examined in a meaningful, quantitative way using a novice population?*

The studies documented in this report aim to determine the feasibility of using novices to gather meaningful, quantitative data that generalizes to reactor operators. The research proceeded in two stages described in Section 2.1 Study Overview.

2.1 Study Overview

The first study establishes the approach and provides baseline performance data from systematically testing non-operator participants with various combinations of scenarios, system conditions, and new technologies to identify leading human performance indicators (i.e., aspects prone to error common to all humans). For instance, operators are required to monitor many plant parameters simultaneously, a novice population can be used as a surrogate to understand what types of displays might cause more monitoring errors or establish guidance for limits on the number of parameters that can be simultaneously monitored. This kind of "different by equal approach" can support researchers' identification of safety-critical or error-prone operational contexts, which could be further studied with operators if needed.

The second study in this report validates the findings of study 1 to determine if results from novices can be reasonably generalized to trained operators. In this a limited number of trained operators were tested in the same error-prone contexts used in study 1. The findings from the two studies will be used to develop guidance for the NRC about potential human factors issues in the operational contexts evaluated in the laboratory.

Work has been done in the NPP domain to understand the types of tasks operators perform (Kirwan & Ainsworth, 1992), but systematic investigations examining performance, the factors contributing to errors, and drivers of operator state have been limited. The present chapter outlines the methodological decisions made by the research team to shape the direction for the research program.

2.2 Proof of Concept

Unlike basic research programs that generate knowledge for knowledges' sake, the applied research of the HPTF must be directly relevant to the concerns of the NRC with a clear application path. To accomplish this relevance, the HPTF aims to:

- Create a cognitively similar environment to an NPP MCR with enough fidelity that the cognitive processes engaged by participants are comparable to those in real operators.
- Demonstrate that novices can successfully perform realistic operator tasks within the proof-of-concept environment.

2.3 Considerations for Conducting Research in the NPP MCR Domain

There are two main types of NPPs operating commercially in the United States: boiling water reactors (BWR) and pressurized water reactors (PWR). About a third of the 93 commercial reactors are BWR (31 of 93). A BWR heats water to steam, which directly powers the turbine. The remaining two-thirds (62) are PWR. In a PWR, water is heated under high pressure to just below the boiling point within the reactor. The pressurized water is then circulated around a lower pressure water system enabling heat transfer to the lower pressure water which turns to steam and powers the turbine.

Commercial reactors tend to use the same technologies as military and research reactors, however, there are some minor differences in the norms for some control room indicators, such as light box and status indicators. For example, in commercial NPPs, a red light indicates a valve or switch is open while in the military domain, red indicates the valve is closed. Specific studies on the impact of congruency between light color and switch/valve status on operator workload and situational awareness are not available in the nuclear domain; it is possible that these types of differences between plant types, room, panel, alert, indicator, and interaction designs, create a fleet of nuclear power plants with unique HSI characteristics.

To fully characterize the workload associated with common NPP MCP tasks, large numbers of ROs would be required to participate in hours of controlled simulation experiments, however, because of limited access to licensed operator and simulator time this kind of rigorous and systematic investigation is not feasible (Hughes et al., 2017; Leis et al., 2014). These limitations have led to research that is primarily qualitative. Common methods used in the nuclear domain include Subject Matter Experts (SME) opinions, industry questionnaires, and small sample studies with ROs. These research methods are good for uncovering directional insights, and guiding future research; however, they lack the sufficient rigor to support theory development and make the necessary statistical inference to support sound regulatory decision making or support system validation (Ha, Seong, Lee, & Hong, 2007).

As part of the preliminary work for the present series of experiments, a literature review on workload in the nuclear domain was conducted. Generally, the small literature was represented by studies with significant design limitations which impacted the applicability of the results to NPP MCR tasks of interest. Moreover, these studies also tended to compare different operating procedures (e.g., Lin et al., 2010) or have other issues such as small sample sizes or using secondary rather than primary task performance for analysis (see Hwang et al., 2008 for example), Mercado (2014) provides a detailed summary of the NPP literature and the impact of these kinds of study design limitations.

The insights drawn from the literature review led the research team to propose an approach that experimentally tested the common task types during MCR operations, balanced task types with all I&C types used and included a large sample of participants to improve statistical inference. While this pilot design was sound and more generalizable than the other NPP MCR workload research, the design utilized novice participants. Hence, it is not yet known if the study findings generalized to operational environments with highly trained ROs.

2.3.1 Nuclear Power Plant Main Control Room Operations

The reactor operator (RO) interacts directly with the HSI of a NPP MCR. The ability to support crew coordination through effective teamwork and efficient communication is vital to MCR

design (Fink et al., 2004). The RO is a member of a NPP MCR crew that has the shared goal of safe and efficient power generation. To achieve safe and efficient operations through well designed systems it is necessary to understand:

- Operator tasks and goals
- Crew communication, coordination, and task execution
- System and interaction design and layout of the MCR

2.3.1.1 Main Control Rooms

All details described herein are with respect to NPP MCRs in the United States licensed to operate under either 10 CFR Part 50 or 52. The MCR houses all the I&C of the reactor and associated safety systems. The MCR boundaries are specified as the “vital area” and are defined in 10 CFR 73.2. MCR controls are defined in 10 CFR 50.54 as any apparatus that directly affects the reactivity or power output levels of the nuclear reactor (U.S. Nuclear Regulatory Commission, 2017). Licensed personnel are required to continually staff the reactor when it is in any operational mode other than refueling or shutdown. Per-shift on-site staffing of an MCR crew is dependent on the number of reactor units and control rooms at the NPP site and is defined in 10 CFR 55.54(2)(i).

A minimum shift complement of NPP crew is made up of at least one licensed Senior Reactor Operator (SRO) in the control room, and one licensed RO present at the controls at all times and one or more relief operators licensed and able to take on the role of operator at the controls. The SRO assigned to “control room duties” is required to be within eyesight or audible range of the operators at the controls (U.S. Nuclear Regulatory Commission, 2008). SRO duties are supervisory in nature. SROs are stationed in the MCR where they have direct and prompt access to information on the current state of the plant. The SRO should maintain situational awareness of the plant’s state, provide expertise and knowledge in the event of an off-normal condition occurring, and execute emergency procedures in the event of multiple alarms or a reactor trip. The RO is primarily tasked with ensuring that the reactor unit is operating safely. The RO at the controls is required to stay within the surveillance area of the MCR with an unobstructed view of the operational control panels and annunciators. Relief operators are also licensed ROs typically tasked with aiding the operator at the controls (U.S. Nuclear Regulatory Commission, 2008).

2.3.1.2 Layout

An efficient, reliable, and consistent HSI is the ergonomic goal for control room design (Raeisi et al., 2016). NPP MCR designs have been developed and modified over many years; as a result, each plant has MCR I&C layouts and workstation configurations that are somewhat unique. Workstation characteristics may differentially impact the functionality of the HSI, and this has not been directly investigated to date. Across the 95 operational NPPs in the United States MCR workstations span the spectrum from fully analog I&C to fully digital (Harris et al., 2017). These kinds of HSI differences not only impact individual crewmember tasking, but also teamwork and cohesion (Fink et al., 2004). Understanding and assessing the impact the HSI has on crewmember tasking is important for an effective MCR design. One approach to understanding the impact of a plant’s unique HSI on crewmember performance is through the requirement that all U.S. operating NPPs have on-site simulators that are sufficiently matched in physical and functional fidelity (i.e., full-scope and full-scale) (Joe & Boring, 2017; Reinerman-Jones, Lin, et al., 2019; U.S. Nuclear Regulatory Commission, 2017). MCR designs leverage the principles in the International Standards Organization’s (ISO) ergonomic design of control centers

(International Organization for Standardization, 2013). ISO 11064-4 specifies recommendations to follow in the ergonomic design of workstations in domains that focus on process control and security. NUREG-0700, in sections 11 and 12, includes specific details regarding a proper workstation and control room configuration.

A MCR contains workstations and other equipment (e.g., spare parts, tools, emergency equipment such as protective clothing, etc.), as well as documentation (e.g., safety procedures and manuals). Ergonomic configuration of the control room is determined by the arrangement of workstations, proper storage and location of equipment, and the organization of document storage for ease of access (O'Hara et al., 2002). MCR workstations are where ROs perform their tasking. They contain HSI elements that control normal operations and the associated plant safety systems. Workstation types vary, but traditionally include, standup-consoles, sit-down consoles, sit-stand workstations, and vertical panels (O'Hara et al., 2002). Ergonomic factors such as control location, visual layout, and overall comfort can all affect RO performance and workload.

2.3.1.3 Communications

The SROs and ROs work as a team with the common goal of operating the NPP in a way that maintains safe and efficient power generation. As teamwork is essential for effective and safe nuclear operations, the SROs and ROs in the same team typically share the same work, training, and rest schedules (Joe & Boring, 2017) to promote coordination and communication within the team. Poor communication has been regarded as one of the main causes of team coordination issues and can negatively affect quality of team and task performance (Kim et al., 2010). Billings and Cheaney (1981) analyzed 28,000 incident reports in NASA's Aviation Safety Reporting System and found that voice communication issues were present in over 70% of the incidents.

The goal for communication during the tasking of ROs is to convey information accurately so that there is a common understanding among the team members. As this communication often involves technical information related to the safety of the plant, it is paramount that all crewmembers understand the information correctly (Min et al., 2004). Most of the communication among team members is in the form of voice communication. While voice communication allows quicker transfer of information, compared to written communication, it is also more susceptible to misunderstanding and other types of errors or inaccuracies.

In order to facilitate effective voice communication, operators in U.S. NPP MCRs utilize a repeat back method known as *three-way communications* (U.S. Department of Energy, 2009). Three-way communication helps ensure the reliable and accurate transfer of information between two people.

A typical three-way message starts with the first communication being a crewmember addressing another by name and issuing a short instruction. The second part of the three-way message has the addressee of the message echoing back the instruction that was understood by the addressee in a paraphrased manner. The paraphrased instruction must contain the technical details of the instruction. Should the addressee need clarification, more detail, or did not understand the instruction, rather than echoing back the message, the addressee would request the needed information via a repeat request. Once the addressee has correctly echoed back the instruction, the initiating crewmember closes the loop by affirming that the instruction

was understood correctly. Kim et al., (2010) found a positive correlation between the use of three-way communication and performance.

2.3.1.4 Tasks in NPP control rooms

NPP's two primary goals are to produce power: (1) safely and (2) efficiently. Both machines and humans in NPPs serve these goals by accomplishing specific tasks. According to the task analysis of NPP Control Room Crews completed for the NRC (NUREG/CR-3371, 1983), operators perform 3 primary functions:

- Supervising and controlling plant operations
- Maintaining plant systems and equipment
- Coordinating plant support activities.

For each primary function, there are a variety of activities or *sub-functions* that support the primary functions (See Figure 2-1). The sub-function are supported by individual *tasks* performed by the control room crews. A task is defined as a set of human behaviors necessary to accomplish a system goal (i.e., sub-function). For example, a subfunction such as "equipment maintenance" may require a series of tasks. For example, it may require stopping the pump (task 1), instructing plant personnel to make some repairs (task 2) and then restarting the pump (task 3). These tasks may need to be performed in a specific order, known as an *operating sequence*. This is the blueprint that designates which tasks are to be performed and in what order for a particular sub-function serving a specific function and goal.

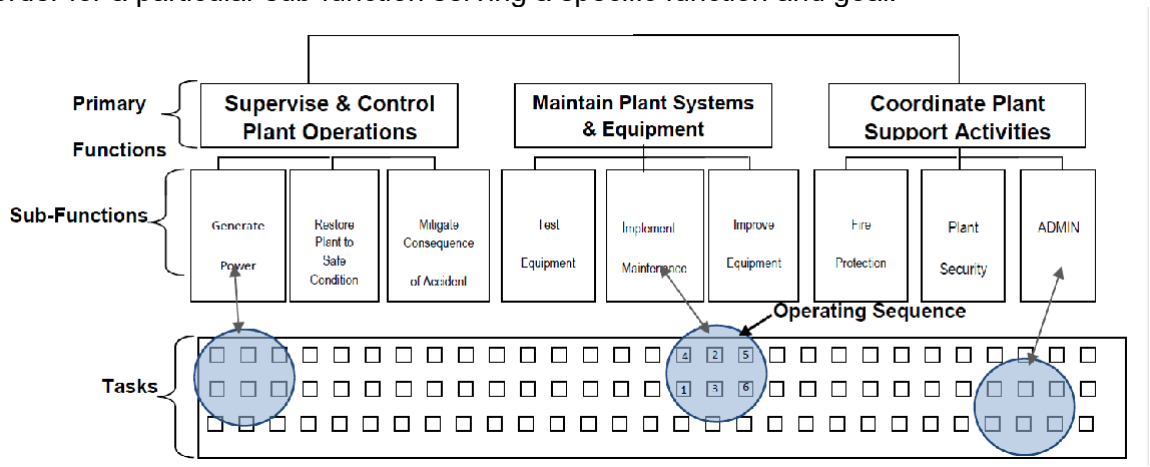


Figure 2-1 Hierarchical representation of operator's role (NUREG/CR-3371).

O'Hara and Higgins (2010) developed a task classification framework that describes primary and secondary RO tasks. Secondary tasks deal with interface management, including navigating, arranging, locating, or accessing information at a workstation. These are considered secondary because they are not directly associated with monitoring or controlling the plant (O'Hara & Higgins, 2010). Understanding secondary tasking is important in the context of primary task performance because the success of the primary task is dependent on the efficacy of information retrieval in the secondary task.

Primary tasks are associated with controlling and running the NPP and include monitoring plant parameters, following defined procedures, as well as manipulating controls to change the state of the NPP. O'Hara and Higgins (2010) defined four distinct elements of primary tasks:

monitoring and detection, assessing situations, planning responses, and response implementation:

> "Monitoring and detection refer to the activities involved in extracting information from the environment. Monitoring is checking the state of the plant to determine whether it is operating correctly, including verifying parameters indicated on the control panels, keeping track of the data displayed on a computer screen, obtaining verbal reports from other personnel, and sending operators to areas of the plant to observe equipment. Situation assessment is evaluating current conditions to confirm that they are acceptable or to determine the underlying causes of any abnormalities. Response planning refers to deciding upon a course of action to resolve the current situation. In an NPP, procedures usually aid response planning; when they are judged appropriate to the current situation, the need to generate a response plan in real time largely may be eliminated. However, even with good procedures, some aspects of response planning will be undertaken. Response implementation is performing the actions specified by response planning. They include selecting a control, providing input to the control, and monitoring the responses of the system and process." (O'Hara & Higgins, 2010, pp. 19-20)

Reinerman-Jones and colleagues (Reinerman-Jones et al., 2013) presented classifications similar to O'Hara and Higgins, however, they regarded monitoring and detection as two separate task types, i.e., the checking and detection tasks. The *checking task* involves a discrete one-time assessment of an I&C to verify its current state or level (e.g., verify that a certain valve is open). It requires observing readings on the displays, viewing I&Cs as well as processing verbal reports from other team members. The checking task type is a successive-attention task where multiple checks are performed back-to-back. The successive-attention component of the checking task maintains consistent demands throughout the task and requires operators to retain critical information in their working memory and distinguish an indicator from a non-indicator (Reinerman-Jones et al., 2006).

Detection is a continuous task where an RO is required to monitor state changes as reflected in the I&Cs and to report back when a certain state has been reached (Reinerman-Jones et al., 2013). The development of this task type was informed by signal detection theory. Detection tasks within the signal detection theory framework generally require participants to remain vigilant and to discriminate noise from signals with noise (Tanner Jr. & Swets, 1954). In the context of the NPP experiments presented later, a detection task would involve an operator monitoring a gauge for changes in level (e.g., Pressurizer level). Two factors influence the operator's ability to detect gauge level changes: (1) signal to noise sensitivity and (2) detection bias. Sensitivity refers to the ability of the operator to discriminate the signal from the noise (Wickens et al., 2015), and is influenced by the operator's ability to remain vigilant at detecting changes over a period of time. Bias impacts an operator's preference for erring on over/under reporting changes. Over reporting bias indicates that the operator is cautious. Thus the operator would rather report a change when one does not exist rather than miss any changes.

Planning response tasks deal with optimally controlling the NPP in a safe state. In the NPP domain, planning responses is typically done using symptom-based Operating Procedures (OPs). These include Emergency Operating Procedures (EOPs), maintenance procedures, and daily operations. OPs are step-by-step symptom- or rule-based procedures defining the appropriate actions to perform on the NPP. The goal of OPs is to maintain the NPP in a safe state or bring the plant back to a safe state optimally in the event of an off-normal event.

The response implementation task type requires the RO to take an action or series of actions on the I&Cs to modify the state of the NPP (e.g., shutting or opening a valve). Response implementation can occur through direct wired analog controls and/or digital soft controls found in modernized and future MCRs.

2.4 Creating an Ecologically Valid Environment

To begin to understand the complex sociotechnical system of an NPP MCR, the system must have been systematically decomposed into the smallest meaningful components. The goal of the studies presented in this report was to demonstrate that it is feasible to construct a cognitively similar environment; tasks and load associated with those tasks are comparable to a real MCR, with an environment (e.g., interface, procedures) that is simplified but representatively realistic. The Skill-Rule-Knowledge human behavior classification framework (Rasmussen, 1983) is a model that researchers have used to aid in understanding interactions between the RO and the HSI (Lin et al., 2010) and it was used as a guide in the present study to determine what kinds of NPP tasks could be used with novice participants.

Taken together, the choices made by the research team served to preserve as much ecological validity of the real operating environment, while still maintaining experimental control to allow for systematic measurement. The following section describes the real environment and subsequently outlines the research team's approach to defining the experimental environment to this effect.

2.4.1 Defining the NPP Simulated Environment

Simulators offer a safe and controlled environment for training and experimentation (Ragan et al., 2015). Most operational NPPs in the United States are primarily still using physical (analog) I&Cs in their MCRs and modernizing by incrementally replacing obsolete parts with digital interfaces (Hugo et al., 2017). Representative of many present-day training simulators, analog full-scope, full-scale simulators employ a spatially dedicated, continuously visible layout and have the fidelity to provide all of the physical and underlying thermodynamics in the real system (Hughes et al., 2017, see also Table 2-1). However, developing, maintaining, and managing such full scope/scale simulators is costly. Additionally, gaining access to these training simulators is challenging for a variety of reasons. Furthermore, as parts wear out in traditional operating plants, the analog I&Cs are being replaced by digital interfaces (Reinerman-Jones et al., 2017).

As interface technologies advance, digital simulators with full-scope thermodynamic capabilities that employ hierarchical layouts of I&C have been developed allowing for enhanced versatility and utility in a variety of contexts (Hughes et al., 2017). That is, while all I&C are available to the operator, they are not continuously in view; rather, they may be displayed in a hierarchical manner embedded within workstation displays contingent upon the purpose of the simulator. When used as a part-task simulator, digital simulators become a potential tool for multiple purposes (e.g., experimentation, operator training, assessment of operator competence, HSI evaluation, and usability tests). Two common workstation designs for digital simulators include desktop with mouse click input and touchscreen input.

As NPP reactor technology and control room design have modernized and evolved, so too have the NPP simulator technology and capability. The upgrades in the NPP MCR may introduce new human factors challenges (Joe et al., 2012). The MCR is where all the I&C are housed that

control the reactor and associated safety systems. The MCR boundaries are specified as the "vital area" and are defined in 10 CFR 73.2. MCR controls are defined in 10 CFR 50.54 as any apparatus that directly affects the reactivity or power output levels of the nuclear reactor (U.S. Nuclear Regulatory Commission, 2017). Licensed personnel are required to continually staff the reactor when it is in any operational mode other than refueling or shutdown. Per-shift on-site staffing of an MCR crew is dependent on the number of reactor units and control rooms at the NPP site and are defined in 10 CFR 55.54(2)(i). There are multiple configurations and levels of capability that warrant description. The HPTF experiments described in this report were conducted in a simulated NPP environment. The characteristics of this simulated environment can be described using the five features defined in Table 2-1. See also Appendix A for a more detailed description of the different simulated environments.

Table 2-1. Summary of types of NPP simulated environments (See also Appendix A).

Features	NPP Simulator Types
Scope	a. Full scope simulator – has the capability to simulate all the physical and underlying thermodynamics occurring in the would-be plant
	b. Part task simulator – has the capability to simulate only part of plant behavior
Layout	a. Spatially dedicated[6] – all I&Cs are available and continuously in view to the operator and presented in a fixed location
	b. Hierarchical – all I&Cs are available but not continuously in view; the I&Cs can be displayed in a hierarchical manner embedded within the workstation displays
Interface types	a. Analog – conventional hard panels or bench boards with hard wired analog I&Cs
	b. Digital – computer-based workstations with digital I&Cs
	c. Hybrid[7] – analog hard panels and computer-based workstations
	d. Simulated Analog – digital representation of emulating analog I&C hard panels
Workstation design	a. Sit-down workstations
	b. Stand-up workstations
Control interaction techniques[8]	a. Mouse click input (for digital and hybrid interfaces)
	b. Touch-screen input (for digital and hybrid interfaces)
	c. Manual manipulations of hard-wired controls (for conventional analog interfaces)

[6] This operational definition of spatially dedicated will be used across all HPTF studies and was developed based on expert guidance.

[7] For the purposes of HPTF related studies, unless otherwise specified, hybrid interface refers to digital analog.

[8] Please note, this is not an exhaustive list of possible control interaction techniques. These are just the ones used in the HPTF research program thus far.

MCR designs leverage the principles in the ISO's ergonomic design of control centers (ISO 11064-4; International Organization for Standardization, 2013). ISO 11064-4 specifies recommendations to follow in the ergonomic design of workstations in domains that focus on process control and security. NUREG-0700 sections 11 and 12 include specific details regarding a proper workstation and control room configuration. As interface technologies advance, digital simulators with full-scope thermodynamic capability become available for the industry and researchers (Hughes et al., 2017). Digital simulators are versatile and have the potential to be adapted to support multiple uses, including operator training, experimentation, assessment of operator competence, HSI evaluation, and usability tests. The simulator used in this experiment can be characterized as: A full-scope simulator with hierarchical layout, simulated analog interface that employed both sit-down desktop mouse-click and stand-up touchscreen workstations.

2.4.2 Human-System Interface (HSI)

NPPs are composed of complex systems that are controlled via an HSI located in the MCR. Most operational NPP MCR in the United States are primarily outfitted with an array of physical or analog, I&Cs spatially distributed around the MCR on large panels. Often, the I&C on these control panels are arranged according to plant and system function and components. The ROs and SRO use I&C to monitor and control the plant. (Savchenko et al., 2018). The simulator used to collect these data was a digital representation of a generic analog NPP MCR panel interface (See Figure 2-3 Original Control Panel Used by Operators (left) and Simplified Control Panel for Novice Participants (right)left).

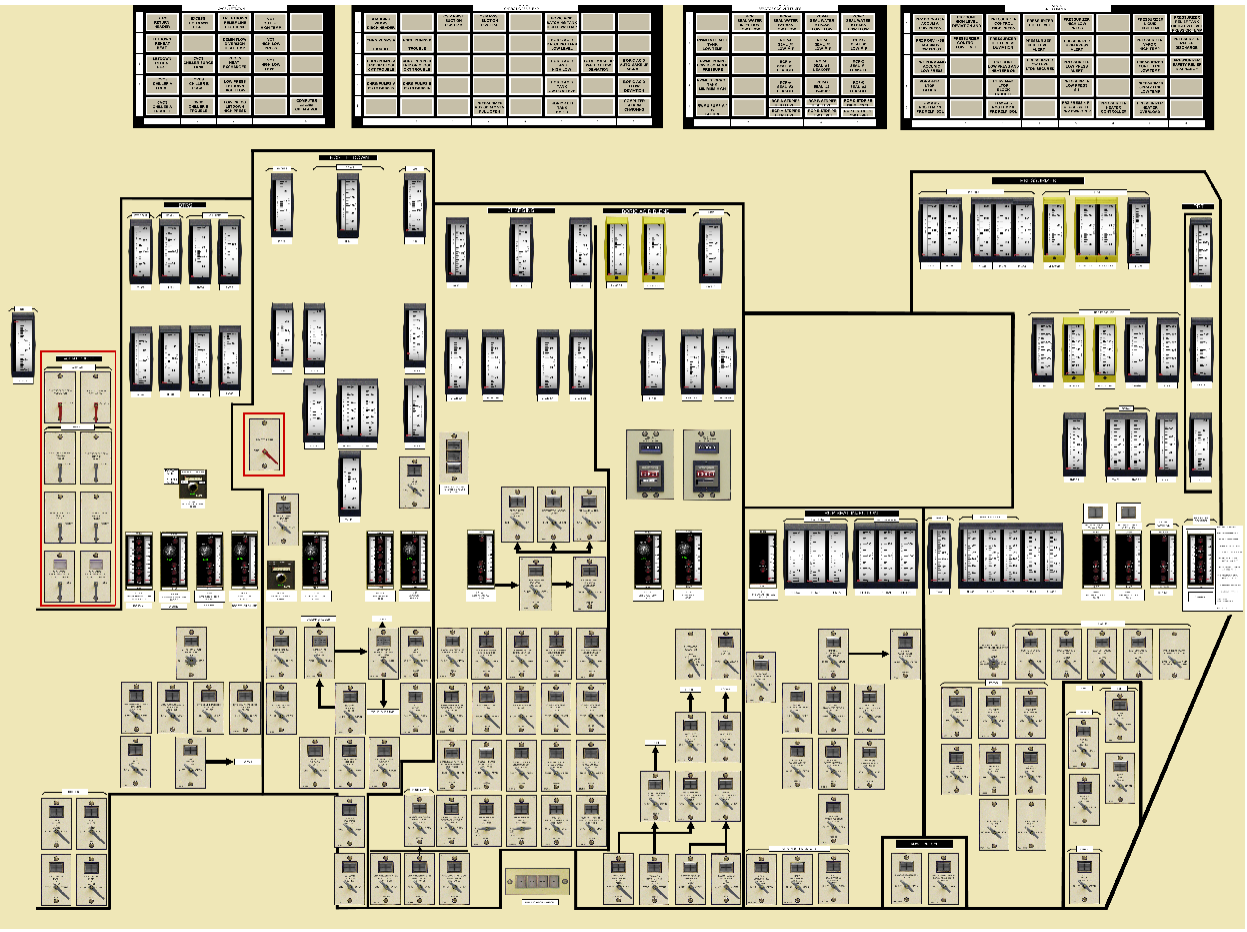


Figure 2-2 An Example of a Digitally Represented Analog NPP MCR Control Panel

2.4.3 Staffing Complement and Conduct

NPP MCRs are managed by teams or "crews" of professional operators. In the current operating fleet in the U.S., a minimum MCR crew is composed of an SRO who directs two ROs to perform steps prescribed in the EOPs to bring the plant to a safe state during emergencies.

The experimental staffing complement included an SRO (played by the experimenter), RO1 (played by a confederate[9]), and RO2 (the study participant). The staffing complement mimics the minimum crew complement required in NPPs, thus, maintaining a team dynamic similar to actual NPP MCRs. In addition, the crew was required to use three-way communication throughout the experiment which is widely used in NPP MCRs and considered an industry best practice.

[9] In experimental psychology research, a confederate is an actor placed into the experiment by the researcher whose role is to play along within the experiment typically unbeknownst to the participant.

2.4.4 Defining the Tasks

We began by first considering all the possible tasks performed by trained NPP MCR operators and determined that, since EOPs are standard across all U.S. control rooms, this would be a good set of tasks to start with. These tasks were used to derive the different but equivalent EOPs to use with the novice participants. Then, several methodological steps were taken to arrive at the three types of NPP MCR tasks that participants would be asked to complete. These tasks were defined based on the earlier work of O'Hara et al., (2008) and Reinerman-Jones et al., (2013) and are checking (monitoring), detection, and response implementation.

The present study focuses on the monitoring, detection, and response implementation MCR tasks. The situation assessment task requires representation of specific domain knowledge and experience. Situational assessment also requires comparing the knowledge domain and experience of the operator with information perceived during observation of the HSI conveying plant parameters. Response planning tasks, like monitoring, detection, and response implementation are largely guided by standardized procedures (e.g., EOPs) and SRO directions.

Table 2-2 Definitions from O'Hara et al., (2008) generic primary MCR operations tasks

Task	O'Hara et al., Definition	HPTF Operational Definition
Monitoring (checking)	checking the plant to determine whether it is functioning properly by verifying parameters indicated on the control panels (see Figure 2-2 An Example of a Digitally Represented Analog NPP MCR Control Panel), observing the readings displayed on screens, and obtaining verbal reports from other personnel.	a one-time inspection of an instrument or control to verify that it was in the state that the EOP calls for it to be (e.g., open or shut). Participants were required to locate various I&Cs by clicking on the correct control. The detection task type required participants to correctly locate a control and
Detection	Perception that the state of the plant has changed.	required participants to correctly locate a control and continuously monitor it for identification of change. Participants were required to monitor the gauge for five minutes and detect changes by clicking on a button located at the bottom of the display.
Situation assessment	evaluating the current state of NPP systems to determine if they are within required parameters.	
Response Planning	deciding upon a course of action to address the plant's current situation. Response planning tasks consist of deciding on a plan to diagnose and perform appropriate actions when an event occurs.	

Response Implementation	performing actions required by response planning (i.e., as directed by the EOP). Response implementation might include selecting a control, performing an action on the control, and watching responses of the system and process resulting from the action (O'Hara et al., 2010).	required participants to correctly locate a control and manipulate it in the required direction. Task type presentation was partially counterbalanced such that the checking always preceded the response implementation because, in a real operating scenario, an operator would never implement a response prior to checking the state of the instrumentation first.

2.4.5 Reducing Overall Complexity While Maintaining Fidelity

One of the aims of the HPTF is to create a cognitively similar environment to an NPP MCR with enough fidelity that the cognitive processes engaged by participants are comparable to those in real operators. To accomplish this some difficulty reduction was required. Difficulty reduction was achieved by creating experimental scenarios that required the use of only two control panels. Next, each panel was modified to reduce overall complexity. Specifically, the panels were modified by reducing the amount of I&Cs contained on each panel and changing the naming convention of the I&C. The original names of the gauges and switches were long alpha-numeric sequences (length > 7 items) that would be meaningful to an experienced operator, but to a novice would appear arbitrary. This distinction means that the memory load for the original labels would be higher for the novices than the experienced operators. To adjust for this difference and reduce the memory load imposed by the labels, each identifying label was modified to reduce its length to around seven characters (see Miller, 1956; Cowan, 2010; Reinerman-Jones et al., 2013, for background on the "magic number" 7 +/- 2). To ensure differences were due to task performance and not differences in the number of I&C items, the two panels were arranged such that the number of items was the same on each panel. This required a systematic reduction in items on each panel. To perform this systematic reduction, the original panel with the fewest number of controls was identified – in this case, panel C1. Next, a systematic reduction of the number of I&Cs on the A2 panel occurred based upon a calculated percentage to equal the number of controls on panel C1, which had 113 I&C elements (see Figure 2-3 Original Control Panel Used by Operators (left) and Simplified Control Panel for Novice Participants (right)).

The I&Cs were categorized into five groups for experimental purposes including:

- gauges
- switches
- breakers
- light boxes
- status boxes

Participants interacted with gauges, switches, and light boxes. Each type of I&C was reduced by the previously calculated percentage, thus leaving the ratio of I&C types the same on each panel. This systematic approach ensured the complexity of the original panel remained. In other words, the ratio of I&C on the modified panel remained intact relative to those of the original panel. For further detail on these modifications, see (Reinerman-Jones et al., 2013).

It is important to note that this modified HSI was not interacting with the reactor thermodynamic physics model. After a series of pilot tests using the modified panels, we determined that having the simulator respond dynamically[10] to operator input did not allow for sufficient experimental control critical to human-in-the-loop experiments. Therefore, we removed the physics model, forgoing the dynamic simulation environment for a controlled experimental environment able to be systematically presented to participants allowing for statistical analysis of their performance. However, the order in which certain steps occurred within each task type, as well as the timing and incremental changes in system parameters (e.g., temperature and pressure) were maintained in accordance with the would-be physics of a dynamic environment experienced by real operators.

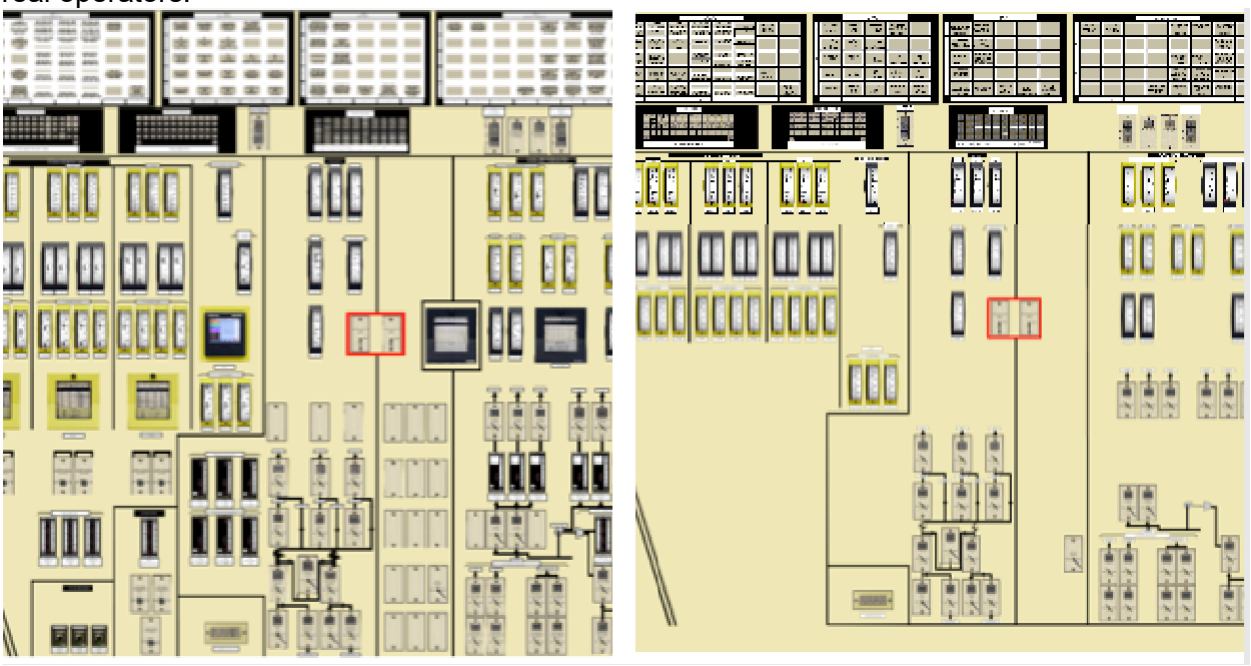


Figure 2-3 Original Control Panel Used by Operators (left) and Simplified Control Panel for Novice Participants (right)

2.5 Measuring Human Performance

Human performance is defined by Gawron (2019) as the accomplishment of a task by an operator or by a team of operators. Tasks can vary from simple (card sorting) to complex (landing an aircraft, reading an x-ray, controlling a nuclear reactor). Humans can perform the task manually or by monitoring an automated system. In every case, human performance can be measured.

The human factors community continually seeks improvement to the tools and methods, also known as measures, used to characterize and measure human performance. As the measures

[10] Dynamic response of simulator refers to the resulting change to the state (i.e., the physics) of the simulator, from a thermodynamic perspective, based on operator input.

used often can be limited by the environment in which they are administered, part of this improvement process is finding measures best suited for the domain of interest. A need exists to identify measures of human performance related to safety in NPPs (Hallbert et al., 2006). One of the aims of this research program was to discover more information about the compilation of measures best suited for the nuclear domain through the identification of measurement tools (i.e., measures) most sensitive to changes within this environmental context.

2.5.1 Using a Training Simulator for Human Performance Experimentation

In the real-world, operator actions change the behavior of the plant and influence the subsequent actions that operator can take. This variation makes it difficult to attribute performance to any one specific task- or operator-related factors. To eliminate this challenge, a GSE Generic Pressurized Water Reactor (GSE GPWR) was modified such that the research participants engaged in operationally realistic tasks, but in an experimenter-controlled order, and with stable simulated plant responses. To accomplish this, JDesigner™, GPWR™, and EPIC, were used to support the development of the experimental scenarios run on the GPWR NPP MCR simulator platform.

- JDesigner™ is an interface design tool. It was used to construct virtual panels for the NPP MCR experiments.
- GPWR™ is a full-scope model of a generic NPP that links the panel's I&Cs to the physics of the simulated PWR.
- Experimental Platform for I&Cs (EPIC) software is custom developed software that mimics the user interface panels from JDesigner™.

The key difference between JDesigner™ and EPIC is that the I&C states on the panel have a limited and pre-defined range of behaviors that is controlled by scripts, rather than the GPWR physics model. This enabled a realistic experience of the HSI that was experimentally controllable and therefore repeatable and consistent across all participants, regardless of their actions on the HSI.

2.6 Understanding Operator Workload during NPP MCR Operations

Performance outcomes are not the only way to assess operator performance. To ensure safe operations it is also important to understand the operator experience as they perform tasks. The operators ongoing experience can be directly tied to performance outcomes, understanding how their experience changes as task demands change can be done by measuring the workload associated with the operators tasks (Cain, 2007).

2.6.1 Defining Workload

Workload can be broadly defined as the mental cost of performing tasks, the construct seeks to answer: "how busy is the operator?" and "will the operator be able to respond to an unexpected event?" (Wickens, 2015). Although there is not a universally agreed upon definition of workload, all proposed definitions have two fundamental themes. First, all consider workload as an active interaction between the operator and their task (Megaw, 2005). Second, all theorize workload as the amount of information processing, mental effort, and/or cognitive resources required for task performance, relative to their capacity (Kahneman, 1973; Moray, 1979; Gopher & Donchin, 1986; Kramer et al., 1987; Eggemeier et al., 1991; Veltman & Gaillard, 1996; Hockey, 1997; Taylor, 2012; Abich IV., 2013). The impact of operator workload on performance is a widely studied area of human factors. A Google Scholar search for the key terms "workload" and "human factors" yielded nearly 900,000 hits and the seminal Wickens (2008) Multiple Resources

and Mental Workload is currently cited in 1852 other papers (see also (Hancock & Meshkati, 1988). Despite this prevalence, workload research specific to the nuclear domain has been limited (see Reinerman-Jones, Hughes Green, D'Agostino, & Matthews, 2019; Reinerman-Jones et al., 2006).

The NRC's Human Factors Engineering Program Review Model, NUREG-0711, Rev 3 (O'Hara et al., 2012) identified workload as one human performance construct to consider in HFE design. Additionally, Derouin and Salaway (2018) stressed the importance of workload by contending that,

> "Assuring that workload estimates are reasonable strengthens the licensees' response capabilities and reliability in the execution of work activities required during events that may range in severity, from anticipated operational occurrences up to and including severe accidents". – Derouin and Salaway (2018, p. 169)

Workload changes as a function of task demands (Wickens et al., 2015). Within the context of the sociotechnical system, upgrades and modernization of I&C and interfaces can impact the way tasks are distributed and performed thus impacting the associated workload. High workload can result in overloading, leading to stress and fatigue, which degrades performance in terms of the humans' abilities to perceive, detect, and respond to changes in the state of the plant in the case of emergencies or unanticipated events (Tran et al., 2007).

2.6.2 Assessing Workload

As with its varied definitions, perspectives on the appropriate measurement techniques for assessing workload are diverse. This makes measure selection and cross study comparisons challenging. While there are many measurement techniques to choose from, few have been validated for use in the nuclear domain (Reinerman-Jones et al., 2015; Spielman & Hill, 2017; Xu et al., 2017). As such, one of the underlying methodological goals of this program of research is to begin to build rigorous and validated workload assessment techniques to close this gap in the literature and provide a strong technical basis for licensee's selection and implementation of workload assessments as part of their HFE programs.

Workload can be assessed using subjective, performance-based, and physiological measures. Subjective assessments of workload are conducted using questionnaires, such as the NASA-TLX.

2.6.2.1 Subjective Techniques

Many subjective assessments of workload interrupt the task or are performed post-hoc. Interrupting the task changes the overall flow of events and perhaps even the demand requirements of the operators. Questionnaire administration in the middle of a scenario could hinder operator performance and potentially increase error when the task is resumed, or conversely, because a "break" allows the operator to reflect on the scenario event thus far, performance could improve. These challenges are avoided with post-hoc measures, however, these might not be sensitive to the dynamic changes occurring in the NPP.

Subjective measures are the most widely used tool to assess workload, likely because of their ease of use and face validity (Estes, 2015). The most commonly administered self-report tool used is the NASA-TLX (Hart & Staveland, 1988). The NASA-TLX is referenced in more than 6000 published works including over 550 reviews of the tool itself. The tool measures six relatively independent subscales: mental, physical, temporal demands, frustration, effort, and

performance. NASA-TLX is most often administered post-task which requires operators to recall events. The ISA is another commonly administered measure. ISA is an online measure developed by the United Kingdom Civil Aviation Authority as a simple, immediate rating of work demand during primary task execution (Tattersall & Foord, 1996). The ISA is administered with a short auditory prompt that signals the operator to rate his/her current global workload on a 5-point Likert scale ranging from being under-utilized to experiencing excessive workload. The present study utilized both the NASA-TLX and ISA as a means to assess the subjective facets of workload.

The potential issues with breaks and the need for a comprehensive approach to workload assessment in the nuclear domain (see Tran and colleagues 2007), led others to also include physiological measures, such as electroencephalogram (EEG), fNIRS, TCD, eye movement, and cardiac indicators.

2.6.2.2 Physiological Techniques

There are many benefits to using physiological metrics as an assessment of mental workload (Matthews & Reinerman-Jones, 2017). Most importantly, physiological metrics provide objective and continuous monitoring of the participant's cognitive and physical state (Reinerman-Jones et al., 2010). Several physiological measures are being considered for inclusion in our NPP test case. Electroencephalogram (EEG) measures neural activity and is sensitive to changes in mental workload. EEG allows for the continuous monitoring of brain activity without interfering with the primary task (Brookings et al., 1996).

However, technical expertise is extensively required for analysis because there are no standardized scoring procedures (Kramer, 1991). This is due in part to variations in physiological response patterns. It has been observed that individuals produce different physiological responses to identical circumstances (Turner, 1994). In addition to individual differences issue, task types themselves produce different patterns (Miyake, 2001). Still, with these limitations, ISO 10075-3:2004 "Ergonomic principles related to mental workload" has recognized the importance of incorporating physiological indices in their workload measurement method (International Organization for Standardization, 2004).

2.6.2.2.1 EEG

EEG has been used in numerous workload studies due to the significant finding that EEG correlates with workload (Berka, Davis, et al., 2007). EEG is a measure of neural activity measured via electrodes placed along the scalp. While event related potentials are common in many low-level cognitive and perceptual tasks, power spectral density analysis has been shown to be sensitive to changes in workload. Power spectral density analysis yields the theta, alpha, and beta frequencies, which have been shown to be sensitive to changes in workload (Berka, Levendowski, et al., 2007; Eggemeier et al., 1991; Hankins & Wilson, 1998; Kurimori & Kakizaki, 1995). Theta, specifically in the frontal lobes and along the midline, during mental concentration tasks, is associated with a high-amplitude (Kubota et al., 2001). Similarly, but in an inverse relationship, alpha tends to decrease as workload demands increase (Gevins et al., 1979). Additionally, alpha has been shown to attenuate as visual scanning task complexity is increased (Gundel & Wilson, 1992). Wertheim's research showed the suppression in alpha during visual scanning tasks was caused by retinal involvement and oculomotor control (Wertheim, 1981). Beta activation has been shown to correlate with cognitive and emotional

processing (Gundel & Wilson, 1992). Specifically, beta has been shown to increase with increases in arousal, attention, and workload (Prinzel et al., 2009).

2.6.2.2.2 Electrocardiogram (ECG)

ECG is a measure of the electrical activity of a heart. ECG is one of the most frequently used physiological measures of workload (Mercado, 2014). HRV and IBI have been used to index workload. The general cardiovascular pattern for increases in workload are characterized by decreases in HRV and IBI (Mulder et al., 2004). Electrocardiography (ECG) measures cardiac activity. HR, HRV, and Inter-beat Interval (IBI) have been found to be associated with mental workload (Jorna, 1993; Kramer, 1991; Roscoe, 1992, 1993; Veltman & Gaillard, 1996; Wilson et al., 1994).

2.6.2.2.3 Transcranial Doppler (TCD)

TCD sonography monitors cerebral blood flow velocity (CBFV) in intracranial arteries and has been commonly used in vigilance studies showing a decrease in CBFV paralleled by decreased performance for sustained attention of highly demanding tasks (Reinerman-Jones et al., 2011). Vigilance is the detection of infrequent signals amidst non-signals or noise. Much of the operators' responsibility fits the criteria of a vigilance task.

2.6.2.2.4 Functional Near-Infrared Spectroscopy (fNIRS)

One way to quantify the extent of mental resource engagement is to measure energy consumption changes in response to task or situational demands. To accommodate the load of ongoing tasks the brain directs resources (e.g., oxygen and glucose) to the pathways associated with cognition (often in prefrontal cortex). Functional Near-Infrared Spectroscopy (fNIRS) imaging is a tool for observing the hemodynamic changes in oxygenated hemoglobin and deoxygenated hemoglobin associated with cognitive activity (see Causse, Chua, Peysakhovich, Del Campo, & Matton, 2017). fNIRS is particularly appealing in human factors and simulation domains because it can be fielded in complex environments and provide real-time or near real-time indexing of cognitive workload. The expected direction of hemodynamic change is that as task difficulty increases blood oxygenation also increases (Ayaz et al., 2010).

2.6.2.3 Performance Measures

Performance measures can also be diagnostic in mapping the relationship between task demands and operator workload. However, performance alone may be a poor indicator of workload because under certain conditions, a dissociation between workload and primary task performance has been observed (Hancock 1995; Leis et al., 2014; Matthews et al., 2015; Mercado, 2014; Yeh & Wickens, 1988).

2.6.2.4 Summary

Most agree that workload is multidimensional and is a result of the demand imposed by the task on an operator's mental resources. However, in the literature there are many conflicting ideas, definitions, and ways to measure workload (Moray, 2013). For the present study, workload is defined as:

> "the operator's perceived evaluation and accompanying physiological response to the experience imposed by the task demands rather than a direct reflection of the task demands themselves" (Abich IV., 2013, p. 223).

Since the 1960s, the measurement of workload has been a significant area of research (Estes, 2015). By the late 1970s, researchers started measuring workload through subjective ratings, expert opinions, performance-based tasks aimed at quantifying spare mental capacity, primary performance tasks, as well as physiological correlations (Williges & Wierwille, 1979). The present study aims to triangulate using multiple measures, which will more accurately reflect the multidimensionality of workload.

2.7 General Methods

As part of our general methodology, we established the following research questions as the focus of our work:

- Can novice participants perform proficiently on realistic operator tasks?
- If novices can perform proficiently, are there differences in performance as a function of task type?
- What is the level of workload associated with various types of tasks?
- What is the type of workload associated with each task type?
- What types of performance errors tend to occur for each task type?

When the research program first began, questions about what should be measured and how it should be measured needed to be answered. The final stage in the process of developing the methodology is selecting measures that allow us to understand performance, determine error types, and understand the state of operators (stressed, overloaded, alert, etc.) while interfacing with complex systems. To accomplish this, objective task performance had to be assessed. Objective task performance can be measured in terms of response time, accuracy of actions, and detection of changes. Errors can be categorized along dimensions of slips, lapses, violations, and mistakes (Reason, 1995).

The present experiments assessed operator workload via three categories of measures identified by Eggemeier et al. (1991): subjective rating scales (self-assessment), and two objective forms of measurement, performance-based and. Subjective measures are typically in the form of questionnaires (Kahneman, 1973; Moray, 1967) or Multiple Resource Theory (MRT) measures (Wickens, 2008; Wickens et al., 2015). Performance measures come in the form of primary and secondary task performance, where decrements indicate a change in workload (Wickens et al., 2015). Physiological measures continuously monitor bodily responses to associated changes in task load (Cain, 2007). By using a comprehensive selection of workload measures, the present research sought to determine which measures might be best suited (i.e., most sensitive and diagnostic) for the nuclear domain in the context of the MCR.

The following sections describe the performance measures for objective task performance, subjective and objective workload that were used for both studies described in this report.

2.8 Experimental Design

This section describes the independent variables (factors to be manipulated or measured) and the dependent variables (the data types collected during the study).

2.8.1 Independent variables

The independent variables in Study 1 were task type (i.e., checking, detection, and response implementation), and interface type (i.e., desktop interface and touchscreen interface). The independent variables in Study 2 were the same task types as Study 1 and the RO role (i.e., RO1, RO2) was added.

2.8.1.1 Task Type

The task type consisted of three conditions.

- The checking task type required a one-time inspection of an I&C to verify that it was in the state that the EOP called for it to be. Participants were required to locate various I&Cs and indicate identification by clicking on the correct I&C.
- The detection task type required participants to correctly locate an instrument then continuously monitor that instrument parameter for identification of change. Participants were required to monitor the instrument for five minutes and detect changes in level by clicking on a button located at the bottom of the instrument. Twelve random changes per minute occurred, totaling 60 changes per detection step.
- The response implementation task type required participants to locate a control and subsequently manipulate the control in the required direction (i.e., open or shut).

Each task type consisted of four steps that were executed using three-way communication led by the experimenter acting as the SRO.

2.8.2 Dependent Variables

2.8.2.1 Performance Measures

Performance measures were captured in terms of execution and communication.

2.8.2.1.1 Execution Performance

Task execution was measured via verifiable actions. Verifiable actions are all interactions with the interface.

Table 2-3 Execution performance responses and variables

Response Type	Variables Captured by Simulator
Checking	Correct and erroneous identification
Detection	Hits, misses, and false alarms
Response implementation	Correct and incorrect actions

2.8.2.1.2 Communication (Instruction) Performance

A Kinect with Microsoft Voice Recorder captured verbal three-way communication. Three-way communication performance measures included instruction events per task, instruction events repeated, instruction clarifications, location help, and percent correct. Instruction events per task were the number of three-way communication events completed. An instruction event repeated was the number of requests by participants for a repeated instruction and the number of

requests by the SRO for a repeated response from participants. An instruction clarification was a clarification by the SRO to a participant. Location help was the number of requests, by participants, for assistance in locating the correct control. Percent correct was the percentage of correct responses, on all six parts of three-way instruction.

2.8.2.2 Subjective Measures

As section 2.6.2.1 indicated, there are as many measurement techniques for workload as there are definitions. One goal of the HPTF is to provide technical bases supporting the development of guidance related to licensees' measurement of operator workload as part of their HFE programs. To support this goal, and the establishment of baseline data to be used in future HPTF research, a few common subjective measures were selected for inclusion in the first two studies. Additionally, as each measure was developed for use in slightly different domains and aims to quantify different workload dimensions, the diversity of measures should provide a strong foundation for guidance and enable well-informed down selection or refinement depending on the scope of follow-on research.

2.8.2.2.1 NASA- TLX

The NASA-TLX (Hart & Staveland, 1988; Hart, 2006) multi-dimensional questionnaire was used to assess each participant's subjective workload. Subscales included mental demand, physical demand, temporal demand, effort, frustration, and performance. The NASA-TLX uses a 100-point sliding scale to rate each subscale. The average score of the six subscales provided a separate measure of global workload. Participants received a copy of the questionnaire with subscale definitions and completed the NASA-TLX at the end of each task type, throughout the scenario.

2.8.2.2.2 ISA

The ISA (Hulbert, 1989; Jordan, 1992) was used to measure immediate subjective workload on a five-point Likert scale during the performance of a task (Tattersall & Foord, 1996). Participants received a copy of the measure with definitions and completed the ISA halfway through each task type using a customized computer program that automatically activated an audio prompt containing the questionnaire. The audio prompt contained the phrase, "please rate your workload," (Study 1), "RO1 [RO2] please rate your workload" (Study 2) signaling participants to respond by writing down their rating on a sheet of paper.

2.8.2.2.3 Multiple Resource Questionnaire (MRQ)

The MRQ was used to characterize the types of the mental processes engaged during each task (Boles & Adair, 2001). The items on the questionnaire were derived from factor analytic studies of lateralized processes (see Boles, 1991, 1992, 1996, 2002 for additional description). Participants received a copy of the scales, with definitions, and completed the MRQ at the end of each task type, throughout the scenario. Boles (1996) indicates that the MRQ is most effective when using targeted scales, that is, rather than running the whole MRQ, researchers should only administer subscales relevant to the task under evaluation. The following 14 of 17 scales were included for the present experiment: auditory emotional process, auditory linguistic process, manual process, short-term memory process, spatial attentive process, spatial categorical process, spatial concentrative process, spatial emergent process, spatial positional

process, spatial quantitative process, visual lexical process, visual phonetic process, visual temporal process, and vocal process.

2.8.2.3 Physiological Measures

2.8.2.3.1 Electroencephalogram (EEG)

The Advanced Brain Monitoring B-Alert X10 system was employed to assess nine-channels of EEG and one channel of ECG (Figure 2-4 ABM's X 10 EEG/ECG system). Following the international standard 10-20 System, the sampling rate of 256 Hz captured signals from Fz, F3, F4, Cz, C3, C4, Pz, P3, and P4. Reference electrodes were placed on each participant's mastoid bone. Power Spectral Density analysis techniques were used to analyze three standard bandwidths: theta (4-8 Hz), alpha (9-13 Hz), and beta 14-30 Hz (Wilson, 2002). Each bandwidth was collected for the nine channels. They were combined to compare left and right hemispheres and frontal, temporal, and parietal lobes.

Figure 2-4 ABM's X 10 EEG/ECG system

2.8.2.3.2 Transcranial Doppler (TCD)

The Spencer Technologies' ST3 Digital Transcranial Doppler, model PMD150, was used to monitor CBFV of the medial cerebral artery in the left and right hemisphere through high pulse repetition frequency (Figure 2-5 Spencer Technologies' ST3 Transcranial Doppler). The Marc 600 head frame set was used to hold the TCD probes in place.

Figure 2-5 Spencer Technologies' ST3 Transcranial Doppler

2.8.2.3.3 Functional Near-Infrared Spectroscopy (fNIRS)

The Covidien Invos Cerebral/Somatic Oximeter, model 5100C, was used to measure (hemodynamic) changes in oxygenated hemoglobin and deoxygenated hemoglobin in the prefrontal cortex of the left and right hemispheres (Ayaz et al., 2011; Chance et al., 1993) Figure 2-6 illustrates the fNIRS waveform data generated during the measurement interval .

Figure 2-6 Functional Near Infra-Red (fNIR) spectroscopy

2.8.2.3.4 Electrocardiogram (ECG)

The Advanced Brain Monitoring System B-Alert X10 system was used to monitor the ECG, sampling at 256 Hz. Single-lead electrodes were placed on the center of the right clavicle and one on the lowest left rib (Figure 2-7 Electrode locations for the ECG system). HR was computed using peak cardiac activity to measure the interval from each beat per second. The "So and Chan" QRS[11] detection method was used to calculate Inter-beat Interval (IBI) and Heart Rate Variability (HRV) (Taylor et al., 2010). This approach maximizes the amplitude of the R-wave (Henelius et al., 2009).

[11] QRS refers to the Q wave, R wave, and S wave, which denote specific wave valence (positive, negative) relative to the order in which the waves appear.

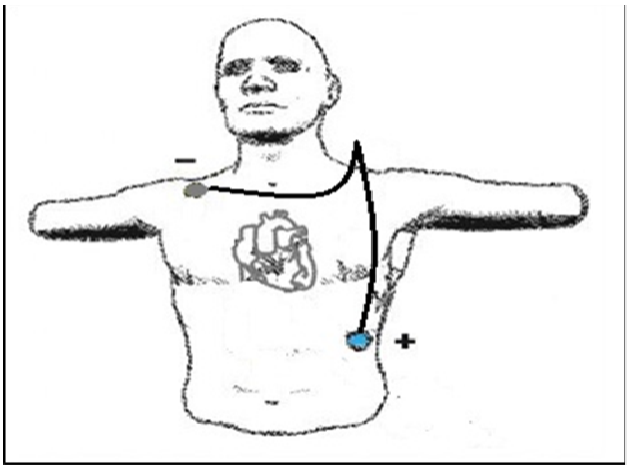


Figure 2-7 Electrode locations for the ECG system

2.9 Summary

Using several of the many possible measures of performance, errors, and states along with understanding the scope and limitations of the operating environment (i.e., simulator capabilities/limitations, physical space, the modified EOPs, required team interaction) will enable identification of measures of workload best suited for particular tasks or a combination of tasks, the levels of workload associated with tasks, and the kind of workload induced (e.g., physical, cognitive) by tasks.

Further, these methods are expected to inform the improvement of data collection techniques for use with the operator population, allowing for additional sources of insight, and generalization of laboratory findings.

3 STUDY 1

The present investigation examines workload levels and types for three common NPP MCR tasks performed on two different interface designs in a controlled experimental environment. The results suggest that the three task types differ in the levels and types of workload imposed. These findings can be used to better understand the types of NPP tasks that induce workload and the type of workload they induce. As part of the methodology development, investigators identified and operationally defined three types of tasks: checking, detection, and response implementation. Task type presentation was partially counterbalanced to maintain ecological validity while imposing experimental control. A variety of subjective and physiological measures were collected to characterize performance on each task in the context of the subjective and physiologically observable workload imposed by task demands. The simulator used to collect these data was a digital representation of a generic analog NPP MCR interface (i.e., simulated analog).

3.1 HSI Modernization

According to the International Atomic Energy Agency, Power Reactor Information System dashboard (https://pris.iaea.org/pris/worldstatistics/underconstructionreactorsbycountry.aspx) there are 57 NPPs are under construction around the world. There are five new Generation III reactor designs each with a different modernized MCR design. These new designs fundamentally change the HSI for the ROs. Many of the 93 U.S. commercially operated NPPs have gone through some partial modernization of the NPP MCR. However, none have completed a full control room modernization effort (Joe, Boring, & Persensky, 2012). Modernizing efforts directly impact the HSI and include changes to alarms, displays, and I&Cs in the MCR (Fink et al., 2004). Digital systems have been adopted in other critical process control domains, but advanced digital I&Cs in the nuclear domain is largely untested (Joe et al., 2012). The same gap in knowledge regarding the modernized control room's impact on workload also exists in hybrid analog-digital control rooms.

3.1.1 Interface Technology: Desktop versus Touchscreen

Most operational NPP MCRs in the United States are primarily outfitted with physical i.e., analog I&Cs. This means that levers, switches, and gauges are all physically present for interaction. Since these analog controls often pre-date any application of usability principals, they are often counterintuitive to use. For example, most BWRs use a red light for an energized or "active" component (pump or generator), or to indicate an open valve for a flow path. The commercial nuclear industry is in a major transition period, moving from legacy analog plants, built in the 1960s and 1970s and licensed to operate into the 2040s (Slater-Thompson, 2014), to new digital power plants like the AP1000, which has been constructed and tested successfully in China (Georgia Power, 2019) and approved by the NRC in July 2022 for operation and fuel load (Patel, 2022).

The interface of the AP1000 digital plants is comprised of desktop monitor displays with interaction via keyboard and mouse there is also one large display for the reactor core at the front of the MCR. Traditional analog plant MCRs are comprised of I&Cs that are organized by system and mapped visually on the panels for each system in the plant. It is well established that the analog MCR layout style is beneficial for mental mapping of system functionality and safety, as well as supporting communication among crew members. These Interfaces with larger displays generally enable more information to be presented, but they also occupy more

space and may require the operator to move around more to view the entire display. On the other hand, some interfaces allow more direct input from the user (e.g., using fingers on a touchscreen), while others utilize translated input devices (e.g., using a mouse or joystick). The study of the workload associated with the NPP MCR tasks should incorporate defining characteristics of current and future interfaces to determine their effects on workload, which can impact performance on the tasks. These kinds of comparisons will ensure preservation of the beneficial features of the analog HSI as the digital systems are integrated into existing and introduced in new NPPs.

Given the potential for HSI changes to impact RO performance, new technologies must be compared to older systems to facilitate a better understanding of any related safety concerns. While one of the main goals of any technology change is to improve the human experience by reducing workload, sometimes, "the tasks required to operate the technology may actually increase workload which may, in turn, degrade human performance" (Aldrich et al., 1989). Understanding how different interface technologies might affect operator performance on the common control room tasks and the associated workload is important to ensuring the continued safe operation.

3.2 Research Questions

3.2.1 Primary Research Questions

Broadly, the goals of this effort were to:

- Determining if novices can perform operator tasks proficiently.
- Understanding operator performance in terms of what operators experience during each operationally relevant task.

To support investigation of the broader research goals, specific research questions were developed for study 1:

- Can novice participants perform proficiently on realistic operator tasks?
- Were there differences in the level of proficiency achieved across the three tasks types (checking, detection, response implementation)?
- What are the workload levels and types associated with various types of tasks?
- What types of errors are associated with various task types?
- What workload measures are more sensitive and diagnostic to which types of tasks?

3.2.2 Supplementary Research Questions

3.2.2.1 Interface Technology

Based on early design knowledge of new reactors, both sit-down desktop displays with keyboard and mouse input along with touchscreen displays requiring tactile input are likely to be used in modernized control rooms. In addition, due to even more recent evidence of the use of both desktop and touchscreen for process monitoring and system control in some cases (Hugo et al., 2017; Ulrich, Boring, & Lew, 2015), we know that both types of digital interfaces will be new and different in the nuclear industry. Current operating plants use an analog interface such that operators interact with physical I&C.

Having access to both the mouse-click technology and touch screen allowed us to make comparisons between the two interfaces and to facilitate a better understanding of any related safety concerns. Thus, we proposed the following supplemental research questions:

- What are the types and levels of workload associated with each interface design?
- Is there an interaction between workload, display design, and task type?

It was expected that levels and types of workload associated with each task type would differ depending upon the soft controls interface implemented.

3.3 Method

3.3.1 Experimental Design Details Related to Interface Technology

For Study 1, participants were assigned to one of two groups, corresponding to the two types of interfaces and completed the same three tasks. Interface type (desktop, touchscreen) was a between-subjects factor (i.e., different participants for each group), eliminating the potential for carry-over effects (exposure to one variable influencing performance on another variable), and thus counterbalancing across interface type was not required. Instead, establishing group equivalence was prioritized to ensure that any differences between groups could be attributed to interface type and not initial differences between the groups. The interfaces fundamentally affected how the tasks were performed as interface differences existed in both control layout and access. The Desktop interface required the participant to scroll and use a zoom feature to access a close-up view of the controls. Since participants were interacting with a desktop configuration, they were seated for the duration of the experiment. The Touchscreen interface was able to display the I&C panel in its entirety (i.e., removing the need for scrolling and zooming), but the large interface required participants to stand and move laterally in order to visually scan and interact with the interface. A between-subjects factor design was employed for this experiment, therefore no carry-over effect existed, ensuring that any differences found between the groups could be attributed to the interface type.

3.3.2 Performance Measures

Task execution (i.e., verifiable actions which include any interactions with the interface) and communication performance, subjective, and physiological measures of workload were used for Study 1 as described in Section 2.4 General Methods.

3.3.3 Participants

Participants included both undergraduate and graduate students from the University of Central Florida (UCF). One hundred and fifty-six participants with ages ranging from 18 to 40 (*Mean (M)* = 20.56, *Standard Deviation (SD)* = 3.45) were recruited using an online participant pool. Participants were required to have normal or corrected-to-normal vision (including not being colorblind) and have no prior experience using an NPP simulator or operating a power plant. They were also required to refrain from ingesting nicotine at least two hours prior to the experiment or alcohol and/or sedative medications at least 24 hours prior to the experiment. Four participants were removed from the sample due to failure to successfully complete the training. The final sample size for this study was 152 (85 males, 67 females). Detail describing the training criteria and procedures are in section 3.3.3 and Appendix C.

3.3.4 Training Participants

Novice participants entered the NPP simulator with little understanding or knowledge of NPP operations. Therefore, it was necessary to devise standard training so that all novices started the experimental recording session with the same comprehension level of NPP operations. Each participant needed to be familiar with key tasks, procedures, and the I&Cs used in the experiment. Participants also needed to learn the 3-way communication protocol used in a real NPP MCR. The training elements all contributed to creating an immersive experience that allowed participants to feel and thus respond as real NPP operators.

Training consisted of three phases using a scaffolding approach. Participants were required to pass a proficiency test for each phase with a score of 80% or greater. They were tested on their abilities in three areas: communication, navigation, and task performance. Participants were allowed a maximum of two attempts to pass each phase of training and only completed a second attempt of a training phase if they did not achieve an 80% or greater on their first attempt. In addition, if participants did not receive a score of 80% or greater on the second attempt of any of the three phases, the researcher classified them as ineligible to participate in the study, and they were dismissed. The following sections describe the development of training for use with a novice population.

3.3.4.1 Participant Training Guide

The Participant Training Guide[12] began with a brief introduction to NPPs and their operation using crews consisting of two ROs and an SRO. Following the introduction, three phases of training were conducted, each building on the previous experience.

- Phase 1: three-way communication protocol (3.3.3.3)
- Phase 2: navigating the simulator (3.3.3.4)
- Phase 3: completing the experimental scenario (3.3.3.5)

3.3.4.2 Phase 1: Three-Way Communication

In this training phase participants were trained in a modified version of the three-way communication technique used in actual NPPs. Three-way communication is a method for relaying information and checking for understanding between team members. In operational NPPs, operators employ three-way communication to reduce mistakes and errors. For our purposes, three-way communication was used to help reducing mistakes and errors as well as to maintain the realism of the task environment and demands. This phase of participant training included PowerPoint slides, diagrams, practice, and a proficiency evaluation.

3.3.4.3 Phase 2: Navigating the Simulator

In this training phase participants were trained to use, recognize, locate, read, and act on instruments and controls on the panels. After familiarization with the panels and tools for moving around the panels, participants practiced three-way communication in reporting readings or actions from the panels. Training included both simplified panels with only a single gauge and

[12] The full Participant Training Guide can be found in Appendix B.

the panels that would be used for the experimental sessions. Participants were trained on the panels they would use during the experiment because operators are extensively trained on the panels in the real world. This phase was complete with PowerPoint Slides, pictures, and a proficiency evaluation.

3.3.4.4 Phase 3: Scenario Completion

This phase focused on teaching participants to follow the steps provided by the EOPs. This phase illustrated the paths followed to identify and isolate problems and was led by the SRO. Therefore, communication was key and following instructions for navigating the panel was a must. PowerPoint slides, figures, and a proficiency evaluation were included in this phase of training.

3.3.5 Use of Confederates

Confederates were used in study 1. The use of confederate researchers in experimentation is a common practice for certain domain-specific psychological testing, specifically social and team psychology. Confederates are generally used when researchers want to (1) elicit specific behavior from the participant that might be difficult to observe naturally, (2) study individual performance during team tasks (3) and/or investigate communication manipulations in real-time (Manusov, 2005). Generally, confederates are individuals working with the lead researchers by performing the role of a participant. The intent is for the confederate to remain unknown as an accomplice in experimentation.

Novice participants served in the role of RO1 while confederates served as RO2. Participants and confederates did not interact, but confederates enabled the realism of NPP MCR crews by preserving communications and presence. Three confederates were extensively trained on the experimental tasks and proper mannerisms and responses as to not alert participants to their position as confederates. The confederates were paired with experimenters who served in the role of SRO for the duration of data collection. Crew composition in NPP MCRs is often stable across shifts, therefore that consistency was adhered to through use of fixed partnering in terms of maintaining the ecological hierarchy. The SRO role is a more senior role, which is the reason a confederate was not used for that role in the present experiment. Experimenters are often seen as a more authoritative figure and, therefore, the cognitive evaluation of the experiment SRO would be similar to that of the actual NPP MCR.

The use of confederates was another aspect of the experimental design that supports the creation of an ecologically valid environment (Leis & Reinerman-Jones, 2015). For Study 1, participants served in the role of RO1 while confederates served as RO2. Confederates were extensively trained on the experimental tasks and proper interactions with the participants. The confederates were paired with experimenters who served in the role of SRO for the duration of the data collection. Crew composition in NPP MCRs is often stable across shifts, therefore, that consistency was adhered to via fixed partnering across data collection sessions. Using a confederate model allowed experimenters to emulate the “team” dynamic experienced by real NPP operators but maintain control over the experience of the participant. The following sections describe the methodology applied for the use of confederates in Study 1.

3.3.5.1 *Good Practices When Using Confederates*

Webster and Sell (2007) stated that to achieve the necessary consistency and control of confederate performance that detailed, specific, and consistent training must be provided. This should include, the role the confederate will play, behaviors that are acceptable for confederates, information that should and should not be said during experimentation, and other confederate requirements (e.g., evaluations administered during training, realistic speech delivery, etc.; Webster & Sell, 2007). To reduce the likelihood of performance errors and confusion, experimenters must provide a clear and concise script and a list of standardized responses for anticipated questions to each confederate. Depending on the setting where confederates will be used, there may be a need to memorize the script and perform it word-for-word. Even with highly scripted language and actions confederates have the potential to exert undue influence over participant performance. To avoid these issues, confederates should be used minimally and selectively.

3.3.5.2 *Defining the Role of Confederates Performing NPP Tasks*

Consistent with Webster and Sell (2007) the introduction of confederates into the study must be done in a targeted fashion, with specific roles and actions defined. To achieve this a task analysis of each role the confederate researcher is to play (Figure 3-1 RO Confederate Task Analysis).

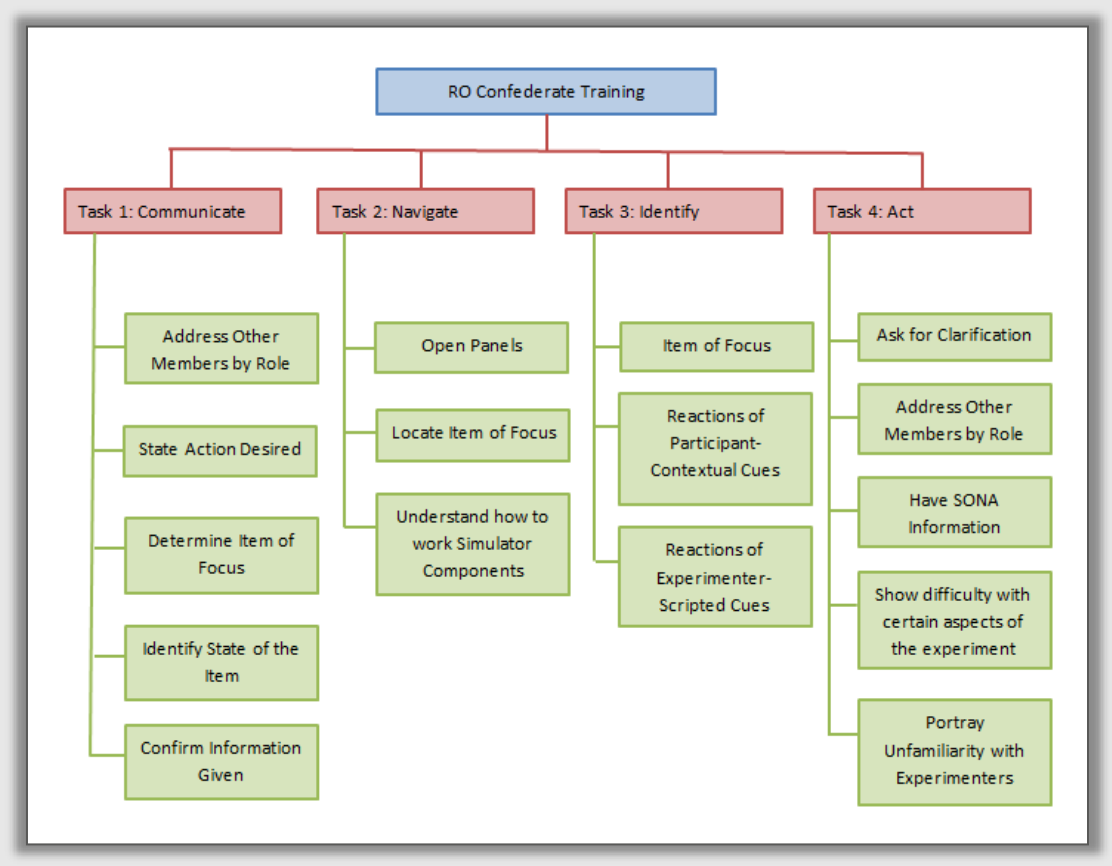


Figure 3-1 RO Confederate Task Analysis

Here, we determined task requirements and explained target behaviors. Armed with this information, we began creating confederate training materials that consisted of a confederate training guide, a confederate training presentation, and confederate training evaluations.

3.3.5.3 Confederate Training Guide

We created a training manual detailing pertinent confederate information for both general confederate use and confederate details specifically identified for NPP tasks during experimentation. Additionally, the training manual covered details concerning confederate evaluation, scheduling information, and additional requirements. The Confederate Training Guide used can be found in Appendix D.

3.3.5.3.1 Confederate Training Presentation

Based on the information given in the confederate training guide, a matching training presentation was developed in Microsoft PowerPoint. This presentation followed the format of the training manual, which was issued to confederate candidates before presenting. Training presentation notes followed a set script presented to confederates in order to provide uniform training to all confederate candidates.

3.3.5.3.2 Confederate Evaluation

After the completion of each section of training, confederate candidates completed a short section evaluation to ensure they met a minimum proficiency standard. This consisted of a few short questions covering the material just learned. Confederates were required to meet minimum standards to continue to the next section of training. Lead researchers established these requirements to ensure all training was consistent and that confederates were able to retain the material learned.

After completion of the training presentation, confederate candidates were paired and scheduled for a video-recorded narrative training sessions to practice the script with a fellow confederate researcher. During these sessions, confederates evaluated his/her own performance, the performance of their partner, and received feedback from lead experimenters. Confederates were required to complete three narrative sessions and were given ample time during work hours to practice the script word-for-word.

Finally, once experimentation began, confederates were required to fill out an after-session report for each session where they noted any observed performance issues. Additionally, lead experimenters also completed an evaluation of the confederate after the completion of each session. Lead experimenters compiled these forms and evaluated them against each other and experimenter logs.

3.3.6 Equipment

3.3.6.1 Simulator

The GSE GPWR simulator was adapted for the present experiment. The simulator included one standard desktop computer (6.4GT/s, Intel Xeon™ 5600 series processor), an x16 multi-display Octal graphics card, eight 27-inch (16:9 aspect ratio) touch screen monitors, and one soundbar speaker.

3.3.6.2 Interfaces

This study included two different interfaces, one desktop (see Figure 3-2 Desktop interface (bottom figure shows zoom of the top figure)) and one touchscreen (see Figure 3-3 Touchscreen interface). Both interfaces used adapted GSE Generic PWR simulator and comprised one standard desktop computer (6.4GT/s, Intel XeonTM 5600 series processor), two 24” (16:10 aspect ratio) UXGA monitors with a total resolution of 3600 by 1200px, and a USB 3-button laser mouse with a scroll-wheel. The interaction design for the desktop interface required participants to use the mouse and scroll-wheel to view all the controls as not all the controls could fit in the display area of the desktop monitors. Participants had to use the mouse to activate the zoom feature (i.e., click on the “+” to zoom in and “-” to zoom out). The touchscreen interface consisted of eight 27” touchscreen WQHD monitor grid (two high by four wide) with a total resolution of 10240 by 2880px and had a touch-based interaction design.

In both set-ups, there was a 104-key Windows keyboard and a soundbar speaker/microphone that were not required for the three NPP tasks but were needed to administer and record responses to the subjective questionnaires, as well as to record the participants’ three-way communications.

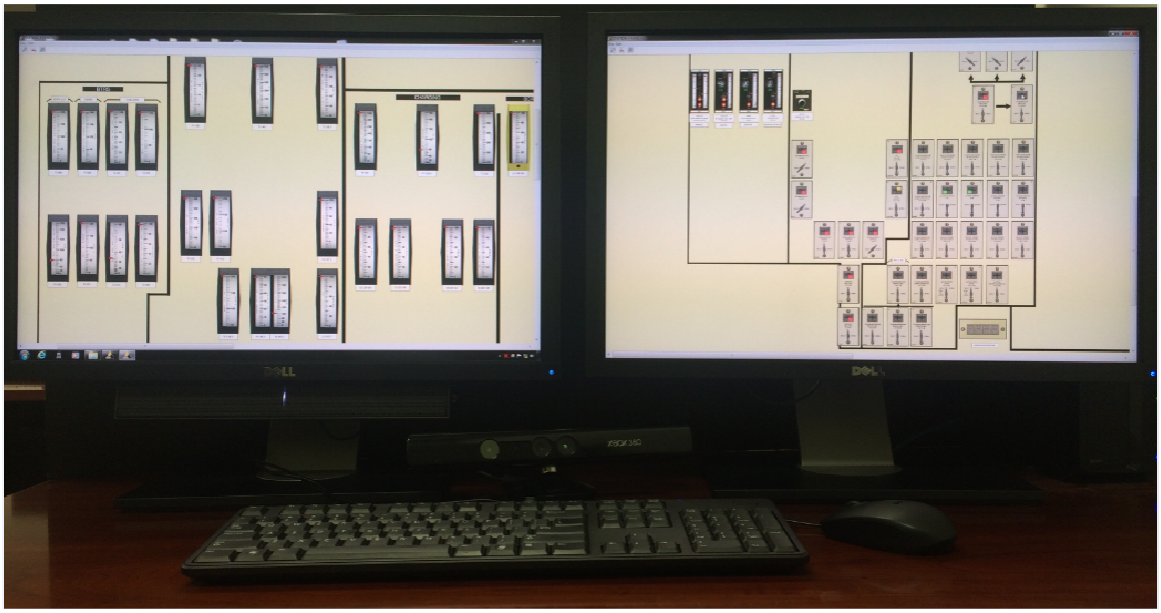


Figure 3-2 Desktop interface (bottom figure shows zoom of the top figure)

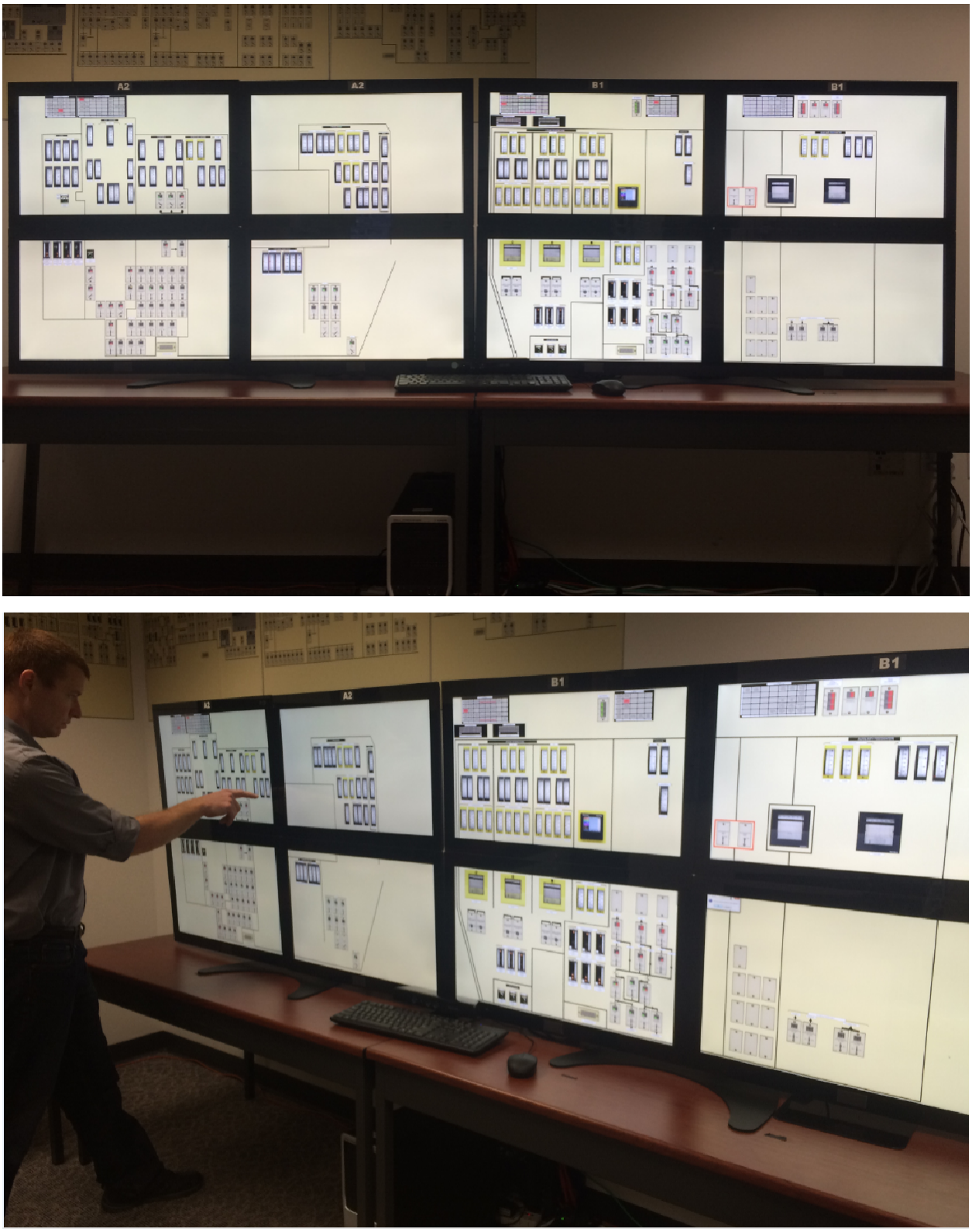


Figure 3-3 Touchscreen interface

3.3.7 Experimental Scenario

The experimental scenario consisted of several types of common control room tasks. The experimental tasks were defined using two approaches: 1) directly extracting tasks from EOPs and 2) through discussions with an NPP MCR operations SME. The initial state of the interface (i.e., I&Cs) was the same for all participants and was derived but decoupled from a valid physics-based scenario. Specifically, the interface reflected the way that I&C would appear if 4a, 4b, 5a, and 5b along with both diesels failed (GSE Power Systems, 2011). This context set the initial simulation parameters. Once established the remaining actions were scripted and not based on the physics underlying the actions taken by the participants.

The experimental scenario required participants to utilize two control panels (C1, A2). To panels were modified so they were appropriate to the skills-base of the novice population Specifically, the amount of controls within each panel were reduced and the naming convention of the I&C was changed (Reinerman-Jones et al., 2013).

The first step to this method was identifying the original panel with the fewest controls to determine the lowest common denominator that could be shared among the panels and retain realism – in this case, panel C1. Next, a systematic reduction of the number of controls on the A2 panel occurred based upon a calculated percentage to equal the amount of controls on panel C1, which had 113 controls. The controls in each panel were categorized into five groups. Gauges, switches, light boxes, circuit breakers, and status boxes. For the present experiment only gauges, switches, and light boxes were used. Each type of control was reduced by the previously calculated percentage, thus leaving the ratio of control types the same on each panel. This approach ensured the complexity of the original panel remained intact. Table 3-1 A2 Panel modification calculation provides the specific modifications to the A2 panel. Figure 3-4 Original A2 panel used by operators (left) and modified A2 for experimentation illustrates the original and modified A2 panels. In addition to enabling a novice population to interact at an appropriate level of complexity, the reduction of the number of controls in panel A2 to equal the number of controls in panel C1 balanced complexity between panels, thereby removing panel as a potential confound.

Table 3-1 A2 Panel modification calculation

I&C type	Number of I&Cs in original panel	Percent reduction needed	Calculated reduction	Number of I&Cs in modified panel
Gauge	108	-43%	61.95	62
Switch	80	-43%	45.89	46
Light box	4	-43%	2.29	2
Status box	0	-43%	0	0
Others	5	-43%	2.87	3
Total	197	-43%	113	113

Figure 3-4 Original A2 panel used by operators (left) and modified A2 for experimentation

The I&Cs are typically represented on MCR panels via acronyms (e.g., MSR BYP SHUT OFF), this means that participants would need to know the acronyms in order to locate the correct control. This type of training was outside the scope of the current study, leading the researchers to adopt a generic naming convention for I&C that contained both an alphanumeric code and name was modified to decrease the complexity of the task environment. Specifically, I&Cs that had an alphanumeric code of greater than seven were recoded to an alphanumeric code of seven or less (i.e., gauge number EI-6963A1 SA was recoded to EI-6963; Figure 3-6 Example of recoding I&C alphanumeric code of greater than seven digits), adhering to Miller's working memory rule of seven plus or minus two items which is the number of items an individual can hold in short term memory (Miller, 1956). Controls that did not originally have a code remained unchanged. Original gauge names, for example, indicated by the red arrow in Figure 3-5 Example of I&C name and alphanumeric code, were not used. Instead, alphanumeric codes, indicated by the grey arrow in Figure 3-5 Example of I&C name and alphanumeric code, were used. Although the names were not removed from the control panel, participants and experimenters were required to only refer to I&Cs by the modified alphanumeric code (i.e., STM HEADER PRESS gauge was gauge PI-464A1).

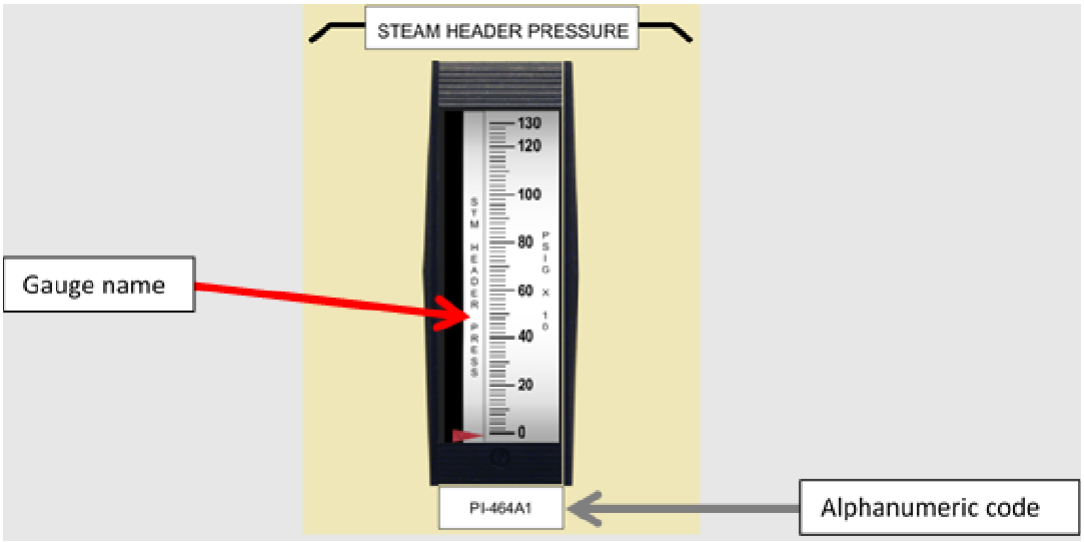


Figure 3-5 Example of I&C name and alphanumeric code

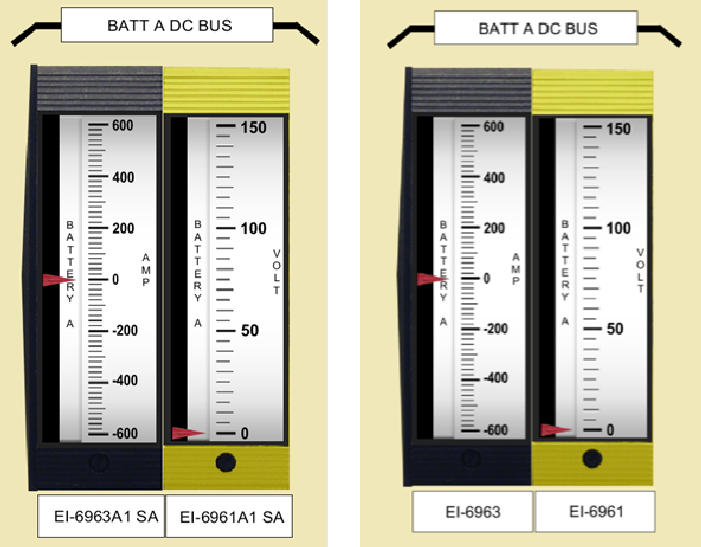


Figure 3-6 Example of recoding I&C alphanumeric code of greater than seven digits

3.3.8 Experimental Design

Data were analyzed using a 3 (task type: checking, detection, and response implementation) × 2 (interface type: desktop and touchscreen) mixed analysis of variance (ANOVA). Task type was a within-subjects factor while interface type was a between-subjects factor. There were twelve steps in each experimental scenario, comprised of one block of each task type (i.e., 4 checking steps, 4 detection steps, and 4 response implementation steps). To limit the possibility of carryover effects, the task types were partially counterbalanced across participants. The task types were only partially counterbalanced because maintaining realism required that the temporally meaningful order of checking occurring before response implementation was maintained (see Table 3-2). Participants were randomly assigned to 1 of 3 scenario orders. Tasks within blocks were not counterbalanced due to practical concerns regarding the necessary sample size.

Table 3-2 Partial counterbalanced task types for scenario generation

	Task Order		
Scenario 1	Checking	Response implementation	Detection
Scenario 2	Detection	Checking	Response implementation
Scenario 3	Checking	Detection	Response implementation

3.3.8.1 Independent Variables

The independent variables in this experiment were task type (i.e., checking, detection, and response implementation), and interface type (i.e., desktop and touchscreen interface).

3.3.8.1.1 Task Type

Task type consisted of three conditions, checking, detection, and response implementation. The checking task type required a one-time inspection of an instrument or control to verify that it was

in the desired state defined in the EOP. Participants were required to locate various I&Cs and indicate identification by clicking on the correct control. The detection task type required participants to correctly locate a control then continuously monitor that control parameter for identification of change. Participants were required to monitor the gauge for five minutes and detect changes in level by clicking on an acknowledge button located at the bottom of the gauge. Twelve random changes per minute occurred, totaling 60 changes per detection task. The response implementation task type required participants to locate a control and subsequently manipulate it in the required direction (i.e., open or shut). Each task type consisted of four steps that were executed using three-way communication led by the experimenter acting as the SRO.

3.3.8.1.2 Interface types

Two types of interfaces were examined: the desktop interface and the touchscreen interface. Participants were assigned to one of the two groups, corresponding to the two types of interfaces. The first group performed the three tasks on the desktop interface while the second group was administered the same three tasks on the touchscreen interface. The facility only permitted one interface to be set up for experimentation at a time, and so after experimentation with the desktop interface was completed, the facility was reconfigured and set up for experimentation with the touchscreen interface. As interface type was a between-subjects factor (i.e., different participants for each group), there was no risk of carry-over effect, so counterbalancing was not required.

There were some differences in the interaction design between the two interface types. The desktop interface required the participant to scroll and use a zoom feature to access a close-up view of the controls. The touchscreen interface allowed all controls to be viewed in their entirety (i.e., removing the need for scrolling and zooming), but the large interface required participants to stand and move laterally to visually scan and interact with the interface, while the desktop interface allowed participants to remain seated.

The detection task required users to engage with the controls every time a change occurred. For the desktop interface, this entailed moving the mouse to click on an area on the control to register detection, but for the touchscreen interface, users merely had to touch the area with their finger. Errors were defined as instances where participants clicked on the area next to the control, i.e., the background, instead of the control.

For the response implementation task, participants were required to manipulate controls to implement an instruction. With the desktop interface, users had to click the edge of the valve “handle”, drag it to position and release the mouse button, whereas with the touchscreen interface, they “touched, dragged, and released their finger” to open/close the valve. These differences may affect the ease with which controls are manipulated, which would be reflected in the number of unsuccessful manipulations or the number of repeated attempts.

3.3.9 Procedure

Participants were provided with a copy of the informed consent, followed by the Ishihara color-blindness test. Participants were informed that another participant (the confederate) who had been trained in a previous session would return for the experimental session at about the time they would complete the training. Participants were then trained for two hours using a PowerPoint presentation and the adapted simulator, on either the desktop or touchscreen

interface. The presentation introduced the procedures and protocols for participating in an NPP simulator for experimental research. Participants were trained to use three-way communication to clearly relay critical information, navigate within the adapted simulator to locate and read status indicators, respond appropriately to a simulated NPP system warning by following standardized procedures, and complete questionnaires. Each aspect was trained separately and then a practice session combined all components. Feedback and proficiency tests were given after each portion. Participants' scores had to be over 80% to move forward to the experimental scenario (see Appendix B and C for detailed information about participant training). After training, participants were given a five-minute break at which point the confederate arrived for the experimental session. The physiological sensors were connected, and a five-minute resting baseline was taken before proceeding with the first task type of the experimental scenario. The steps within the task type were carried out by implementing three-way communication protocol initiated by the experimenter acting as the SRO. The ISA rating was prompted halfway through the task condition block and the NASA-TLX and MRQ were administered after each task condition block. The same process was followed for the next two task type conditions. The experimental session finished by disconnecting the physiological sensors. Experimental sessions were two hours in duration.

3.4 Results

Examination of the group demographics (e.g., gender, age, college major) for the two interface type groups did not reveal any indication that the groups were not equivalent (p[13] > 0.05 for all[14]), therefore demographic factors are not discussed further15. The remaining discussion focuses on training proficiency results, interface and task type effects, and the differences among the subjective and objective measures of performance and workload.

3.4.1 Training

Training consisted of three phases, in which participants were required to pass a proficiency test for each phase with a score of 80% or greater. Participants were allowed a maximum of two attempts to pass each phase of training and only completed a second attempt of a training phase if they did not achieve an 80% or greater on their first attempt. In addition, if participants

---

[13] The "p value" or "p" is used in null-hypothesis significance testing. A null hypothesis is the "default hypothesis" and assumes that two samples are the same. Before a study begins the researcher sets an alpha level to (.05). The alpha level is the probability that a significant result is observed based on chance alone. During analysis, if p < alpha, this means that the two samples are different and the null hypothesis is rejected and that there is a low probability that the null hypothesis was falsely rejected (results produced by random chance). See https://en.wikipedia.org/wiki/P-value for examples and mathematical descriptions.

[14] For Age, $t(123.4) = -1.636$, $p = 0.104$, for Gender, $t(151) = -0.233$, $p = 0.816$, for Major, $\chi^2(5) = 10.339$, $p = 0.066$. These results indicated that the two groups did not differ in terms of the age, gender, and distribution of majors.

[15] The analysis of the demographics (reported in footnote 13) utilized a t test and a χ^2 test. The t test used was an independent samples t test. The independent samples t test compares the means of two independent groups ("samples") to determine if there is a statistically significant difference (indicated by p < .05) between the two group means. A χ^2 is used to examine differences between categorical variables. In this study, age was collected using categorical age ranges thus, χ^2 was used instead of the t test.

did not receive a score of 80% or greater on the second attempt of any of the three phases, the researcher classified them as ineligible to participate in the study, and they were dismissed. A summary of the training performance for the two interface groups is described below and can be found in Appendix C. In general, training performance of the two interface groups for all phases were roughly comparable, although the scores of the touchscreen interface group were slightly lower by 1 to 2 percentage points compared to that of the desktop interface group. The training method utilized seemed effective for both studies with the two different interface groups as indicated by the high scores (above 95%) and low drop-out rates (less than 3%) for all phases.

3.4.1.1 Desktop interface

In the desktop interface group, eighty-three participants attempted Phase 1 training. Four participants failed on their first attempt. Of those four participants, one participant failed to reach proficiency on their second attempt and was dismissed by the researcher. The mean score for the first attempt or second attempts (if applicable) of the 83 participants that attempted Phase 1 training was 95.99% (*SD* = 7.07, *Median* (*Mdn)* = 97.56, *range* = 37.50). For the 82 participants that passed Phase 1 training, the mean score for their first or second attempts (if applicable), was 96.76% (*SD* = 4.91, *Mdn* = 97.56, *range* = 29.27). The participant that did not continue to the next phase of training received a score of 62.50% on their second attempt. The 82 participants that achieved the required level of proficiency during Phase 1 training also achieved the required level of proficiency during their first attempt at Phase 2 training (*M* = 98.63% *SD* = 3.81, *Mdn* = 100, *range* = 18.75). Those same 82 participants attempted Phase 3 training. Of those 82 participants, one participant failed to reach proficiency on the first attempt and second attempt and did not continue with training.

The mean score for the first attempt or second attempts (if applicable) of the 82 participants that attempted Phase 3 training was 98.12% (*SD* = 4.22, *Mdn* = 100, *range* = 25). The participant that did not reach proficiency on their second attempt of Phase 3 training received a score of 75%. Of the 81 participants that passed, their mean score was 98.70% (*SD* = 2.13, *Mdn* = 100, *range* = 7.95). The overall training score (a combination of all three phases) for the 81 participants that reached proficiency and continued to the experimental scenarios was 98.02% (*SD*=3.90, *Mdn* = 100, *range* = 29.27). For those 81 participants, when considering only their passing scores from each phase, their mean was 98.30% (*SD* = 3.01, *Mdn* = 100, *range* = 18.75).

3.4.1.2 Touchscreen interface

In the touchscreen group, seventy-three participants attempted Phase 1 training. Fourteen participants failed on their first attempt. Of those fourteen participants, two participants failed to reach proficiency on their second attempt and were dismissed by the researcher. The mean score for the first attempt or second attempts (if applicable) of the 73 participants (N (number) = 73) that attempted Phase 1 training was 96.02% (*SD* = 8.26, *Mdn* = 97.56, *range* = 68.29). Of those 73 participants, 42.46% (*N* = 31) were female, 57.53% (*N* = 42) were male, and the mean age was 20.15 (*SD* = 2.65, *Mdn* = 19.00, *range* = 13.00). For the 71 participants that passed Phase 1 training, the mean score for their first or second attempts (if applicable), was 97.11% (*SD* = 2.72, *Mdn* = 97.56, *range* = 10.98). Of those 71 participants, 43.66% (*N*=31) were female, 56.33% (*N* = 40) were male, and the mean age was 20.04 (*SD* = 2.46, *Mdn* = 19.00, range = 13.00). The two participants that did not continue onto the next phase of training received a mean score of 57.31% (*SD* = 36.21, *Mdn* = 57.31, *range* = 51.22) on their second attempt. Both participants were males, and their mean age was 24.00 (*SD* = 7.07, *Mdn* = 24.00, *range* =

10.00). Out of the 71 participants that achieved the required level of proficiency during Phase 1 training, 67 achieved the required level of proficiency during their first attempt and 4 achieved the required level of proficiency during their second attempt at Phase 2 training (*M* = 95.68%, *SD* = 5.74, *Mdn* = 100, *range* = 18.75). Out of the 71 participants that achieved proficiency during Phase 2 training, 69 achieved the required level of proficiency during their first attempt and 2 achieved the required level of proficiency during their second attempt at Phase 3 training (*M*= 96.73%, *SD* = 3.16, *Mdn* = 97.72, *range* = 11.36). The overall training score (a combination of all three phases) for the 71 participants that reached proficiency and continued to the experimental scenarios was 96.51% (*SD* = 2.95, *Mdn* = 97.13, *range* = 11.37) (see Appendix B for a summary of training results).

3.4.2 Workload Measures

ANOVA[16]s of subjective and objective metrics were used to determine if there was a significant difference between workload experienced during the three different task types (checking, detection, and response implementation), the two different interface types (desktop and touchscreen), and the interaction between these two factors. Greenhouse-Geisser corrections were used where the assumption of sphericity was not met and, to account for Type I errors, Bonferroni corrections were used for post-hoc comparisons.

3.4.2.1 Subjective Workload Measures

3.4.2.1.1 NASA-TLX

A 3 (task type: checking, detection, and response implementation) × 2 (interface type: desktop and touchscreen) mixed-model ANOVA was conducted for each of the NASA-TLX subscales. Task type was a repeated-measures factor, and the interface type was a between-subjects variable. The ANOVAs were used to determine if there was a significant workload difference between task types, interface types, and if there were overall differences in the ratings across the subscales. The analyses would also reveal if task type effects differed for the two types of interfaces, and if different combinations of task and interfaces elicited different patterns of workload response, as tapped by the NASA-TLX subscales.

A significant main effect was found for task type, $F(2, 296) = 9.663$, $p < .000$, $\eta_p^2 = .061$, such that, in general, participants experienced greater workload during the detection task type (*M* = 38.759) compared to the checking (*M* = 34.302) and response implementation (*M* = 34.035) task types. In addition, a significant main effect was found for the sub-scales of the NASA-TLX, $F(3.067, 453.945) = 50.885$, $p < .000$, $\eta_p^2 = .256$, such that overall, participant reported higher ratings on the Performance (*M* = 47.152) and Mental Demand (*M* = 42.836) subscales compared to the other subscales (Figure 3-7 NASA-TLX scores by subscale (error bars denote standard errors)).

[16] An ANOVA is an analysis of variance and allows for the comparison of more than two groups at the same time. Like the t test described in footnote 14 the ANOVA is used to compare means. The ANOVA produces the F statistic. The ANOVAs reported in this paper are primarily mixed-model, because they contain both within subjects and between subjects variables.

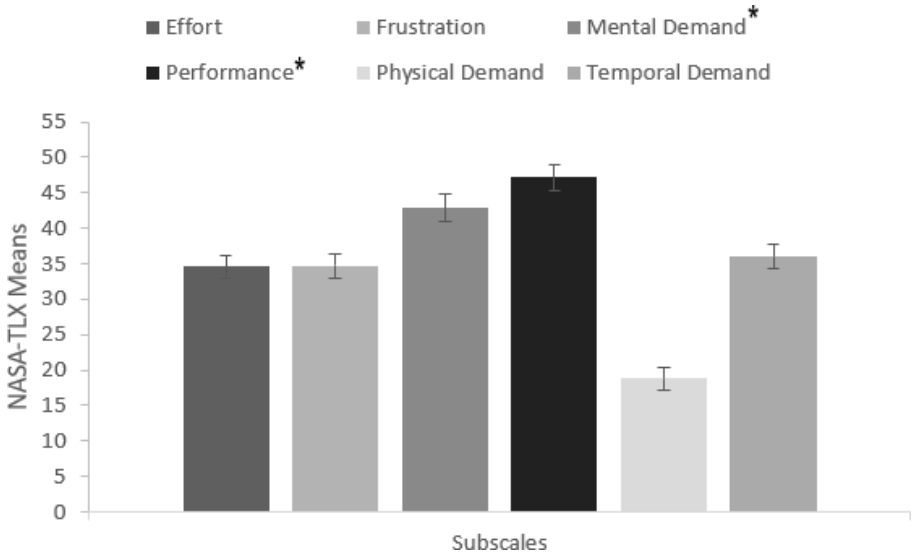


Figure 3-7 NASA-TLX scores by subscale (error bars denote standard errors, asterisks denote significant findings)

Examining the effects of task on the different subscales, results showed a significant interaction effect between the task types and sub-scales on the NASA-TLX, $F(6.705, 992.358) = 19.497$, $p < .000$, $\eta_p^2 = .116$. Not only did the detection task induce the highest amount of workload overall, but it appears that the increase was especially marked for Frustration workload (Figure 3-8 NASA-TLX scores by task type and subscale (error bars denote standard errors)).

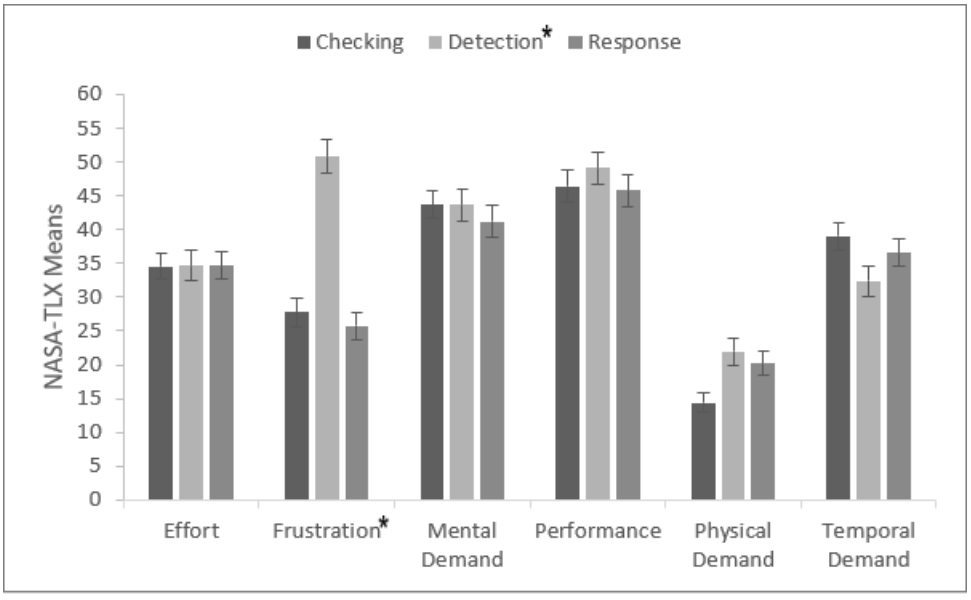


Figure 3-8 NASA-TLX scores by task type and subscale (error bars denote standard errors, asterisks denote significant findings)

Furthermore, a significant main effect for interface type was found, $F(1, 148) = 15.556$, $p < .000$, ${}_p\eta^2 = .095$, such that workload ratings were generally higher for the Desktop interface ($M =$

40.464) compared to touchscreen interface (*M* = 30.934) groups. This increase in workload in the desktop interface group was much greater for the Performance and Effort subscales, as reflected in the significant interaction effect between the sub-scales on the NASA-TLX and interface types, $F(3.067, 453.945) = 21.302$, $p < .000$, $\eta_p^2 = .126$ (Figure 3-9 NASA-TLX score by interface type and subscale (error bars denote standard errors)). The three-way interaction between task type, sub-scale, and interface type was not significant.

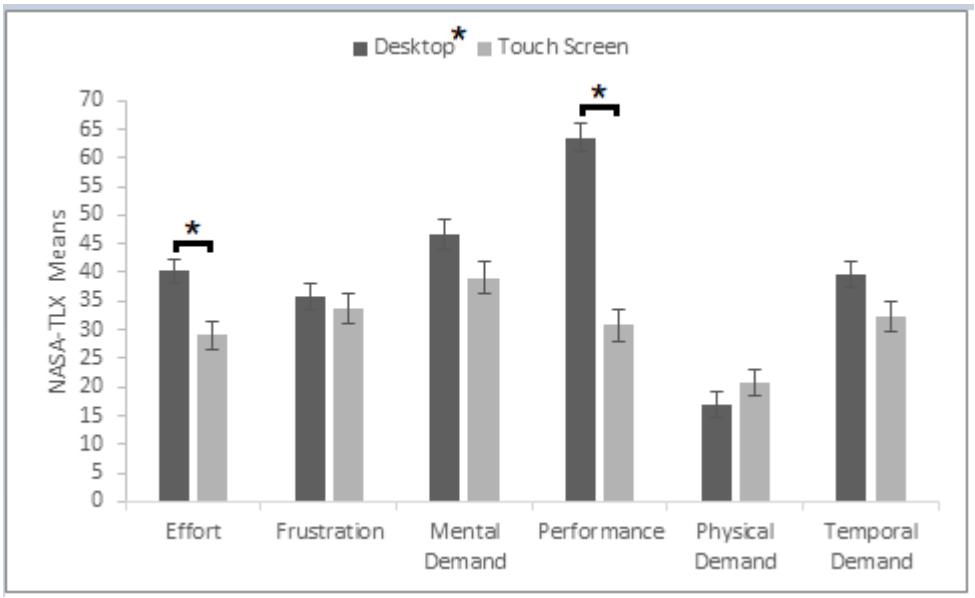


Figure 3-9 NASA-TLX score by interface type and subscale (error bars denote standard errors, asterisks denote significant findings)

3.4.2.1.2 ISA

A 3 (task type: checking, detection, and response implementation) × 2 (interface type: desktop and touchscreen) mixed ANOVA was conducted. As before, task type was a within-subjects factor while interface type was a between-subjects factor. The ANOVA was used to determine if task type and interface type have significant effects on online-subjective workload (i.e., reported as the task were being performed), and to determine if the pattern of workload differences found across the tasks differed between interface types. A significant main effect was found for task type, $F(1.835, 271.627) = 6.149$, $p = .002$, $\eta_p^2 = .040$, such that, regardless of interface type, participants gave higher ratings for checking (*M* = 2.400) compared to the detection (*M* = 2.174) and response implementation (*M* = 2.185) task types. There were no significant differences between the two interface groups in their ISA ratings. The two groups also did not differ on how their ISA ratings differed across the tasks.

3.4.2.1.3 MRQ

A 3 (task type: checking, detection, and response implementation) × 2 (interface type: desktop and touchscreen) mixed ANOVA was conducted for each of the fourteen MRQ subscales tapping various processes that contribute to the workload experienced. The ANOVAs reveal if the task and interface types had any overall effects on the workload, reflected in the activation of the processes as assessed by the MRQ. The analyses would also show whether the effects of task type on the workload ratings were different or consistent across the different types of

interfaces. As before, task type was a within-subjects factor and interface type was the between-subjects factor.

Task type had an overall effect on several of the processes/subscales. A significant main effect of task type was found for the Auditory Emotional subscale, $F(2, 296) = 8.970$, $p < .000$, $\eta_p^2 = .057$, such that the response implementation task ($M = 43.054$) demanded significantly more auditory emotional processing compared to the checking ($M = 34.278$) and detection ($M = 37.182$) tasks.

Task type also had a significant main effect for the Spatial Attentive subscale, $F(1.785, 264.128) = 5.022$, $p = .009$, $\eta_p^2 = .033$, such that the response implementation task ($M = 69.087$) demanded significantly less spatial attentive processing compared to the checking ($M = 72.835$) and detection ($M = 73.977$) task types.

A significant main effect for task type was found for the Spatial Concentrative subscale, $F(1.898, 280.843) = 10.299$, $p < .000$, $\eta_p^2 = .065$, such that the detection task type ($M = 62.372$) demanded significantly more spatial concentrative processing compared to the checking ($M = 52.436$) and response implementation ($M = 57.116$) task types.

There was a significant main effect for task type for the Spatial Positional subscale, $F(1.907, 282.219) = 4.363$, $p = .015$, $\eta_p^2 = .029$, such that the checking task ($M = 70.605$) demanded significantly more spatial positional processing compared to the detection ($M = 66.282$) and response implementation ($M = 66.472$) task types.

A significant main effect for task type for the Spatial Quantitative subscale was found, $F(2, 296) = 18.702$, $p < .000$, $\eta_p^2 = .112$, such that the detection task ($M = 63.779$) demanded significantly more spatial quantitative processing compared to the checking ($M = 51.461$) and response implementation ($M = 52.278$) task types.

Task type also had a significant main effect for the Visual Lexical subscale, $F(2, 296) = 3.825$, $p = .023$, $\eta_p^2 = .025$, such that the checking task ($M = 73.353$) demanded significantly more visual lexical processing compared to the detection ($M = 69.328$) and response implementation ($M = 69.304$) task types.

A significant main effect for task type was found for the Visual Phonetic subscale, $F(1.871, 276.906) = 4.894$, $p = .010$, $\eta_p^2 = .032$, such that the Response Implementation task ($M = 63.830$) demanded significantly more visual phonetic processing compared to the checking ($M = 59.293$) and detection ($M = 58.085$) task types.

There was a significant main effect for task type for the Visual Temporal subscale, $F(2, 296) = 12.142$, $p < .000$, $\eta_p^2 = .076$, such that the detection task ($M = 52.066$) demanded significantly more visual temporal processing compared to the checking ($M = 41.304$) and response implementation ($M = 46.005$) task types.

A significant main effect for task type was found for the Vocal Process subscale, $F(1.818, 268.993) = 14.023$, $p < .000$, $\eta_p^2 = .087$, such that the detection task ($M = 60.835$) demanded significantly less vocal processing compared to the checking ($M = 68.269$) and response implementation ($M = 67.322$) task types.

Interface type also had an overall effect on several of the processes/subscales that cut across all task types. A significant main effect of interface type was observed for the Short Term Memory subscale, $F(1, 148) = 16.588$, $p < .000$, $\eta_p^2 = .101$, such that the desktop interface (M = 82.757) demanded significantly more short term memory processing compared to the touchscreen interface (M = 71.449).

There was a significant main effect for interface type for the Spatial Attentive subscale, $F(1, 148) = 10.113$, $p = .002$, $\eta_p^2 = .064$, such that the desktop interface (M = 76.193) demanded significantly more spatial attentive processing compared to the touchscreen interface (M = 67.739).

Interface type also had a significant main effect for the Spatial Categorical subscale, $F(1, 148) = 4.191$, $p = .042$, $\eta_p^2 = .028$, such that the desktop interface (M = 64.975) demanded significantly more spatial categorical processing compared to the touchscreen interface (M = 59.116).

The effects of Interface type on Spatial Positional processing differed for different tasks. A significant interaction between task type and interface type for the Spatial Positional subscale, $F(1.907, 282.219) = 3.956$, $p = .022$, $\eta_p^2 = .026$, revealed that the greatest differences between Interface types were found during the detection task type (Figure 3-10 MRQ Spatial Positional scores by task type and interface type). No significant main or interaction effects were found for the Auditory Linguistic, Manual, or Spatial Emergent subscales (see Appendix B for a summary of results of subjective measures).

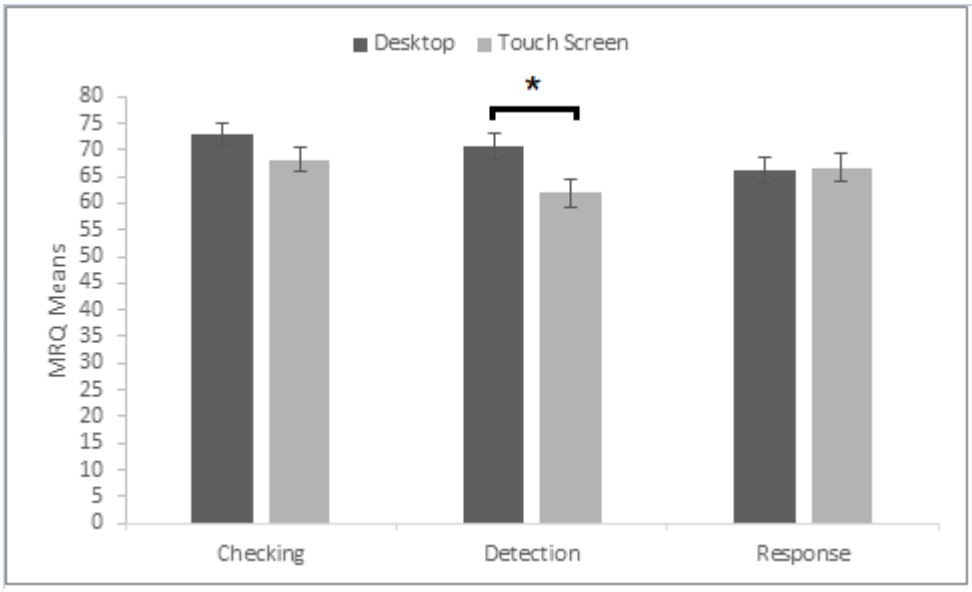


Figure 3-10 MRQ Spatial Positional scores by task type and interface type (error bars denote standard errors, asterisks denote significant findings)

3.4.2.2 Physiological Measures

All physiological measures were derived by taking the difference between a 5-minute resting baseline and the observation interval. For example, if the participant's left CBFV for the five-minute baseline was 73.23 cm/s and their left CBFV for the subsequent checking task was

75.33 cm/s, their difference from baseline would be 2.10 cm/s. This method helps account for individual differences when comparing group means as is the case when running ANOVAs.

3.4.2.2.1 Electroencephalogram (EEG)

Apart from examining the differences in brain activity at all 9 EEG sensor sites, the EEG data could also be analyzed by hemispheres (i.e., compare brain activity between the left and right hemispheres) as well as by lobes (i.e., compare brain activity among the frontal, parietal and occipital lobes). Hence, a series of ANOVAs was performed, as follows:

A 3 (task type: checking, detection, and response implementation) × 2 (hemisphere: left and right hemisphere difference from baseline) × 2 (interface type: desktop and touchscreen) mixed ANOVA, which examined the overall effects of task type, interface type and hemisphere on the Alpha, Beta, and Theta frequency bands. This analysis also revealed if the pattern of interface differences in each of these frequency bands were dissimilar across different combinations of task and hemispheres. Interface type had a significant overall effect on Alpha, $F(1, 148) = 19.320$, $p < .000$, $\eta_p^2 = .115$, Beta, $F(1, 148) = 20.003$, $p < .000$, $\eta^2 = .119$, and Theta, $F(1, 148) = 16.957$, $p < .000$, $\eta_p^2 = .103$, activity. In all cases, the group using the desktop interface showed greater reduction in power within all the frequency bands than in the touchscreen interface group (see Table 3-3 Average change in the various bands from baseline by interface types).

A 3 (task type: checking, detection, and response implementation) × 3 (lobe: frontal, parietal, and occipital lobe difference from baseline) × 2 (interface type: desktop and touchscreen) mixed ANOVA, which examined the overall effects of task type, interface type and lobe on the Alpha, Beta, and Theta frequency bands. This analysis also revealed if the pattern of interface differences in each of these frequency bands were dissimilar across different combinations of task and lobes.

Table 3-3 Average change in the various bands from baseline by interface types

Interface	Avg. change from baseline in Alpha	Avg. change from baseline in Beta	Avg. change from baseline in Theta
Desktop	-34,349.885	-107,618.349	-27,925.687
Touchscreen	299.896	2045.557	1234.206

There was also a significant overall effect of lobe where the Occipital lobe showed greater increases in Alpha, $F(1.650, 244.221) = 9.760$, $p < .000$, $\eta_p^2 = .062$, Beta, $F(1.657, 245.197) = 9.542$, $p < .000$, $\eta_p 2 = .061$, and Theta, $F(1.637, 242.247) = 8.906$, $p = .001$, $\eta_p^2 = .057$ activity from baseline compared to the Frontal and Parietal lobes (see Table 3-4 Average change in the various frequency bands from baseline by lobes). There were no other significant main effects.

Table 3-4 Average change in the various frequency bands from baseline by lobes

Interface	Avg. change from baseline in Alpha	Avg. change from baseline in Beta	Avg. change from baseline in Theta
Frontal	-38,775.713	-116,343.583	-31,075.603
Parietal	-19,610.814	-59,016.101	-14,317.372

Occipital	4563.401	12,596.619	3389.093

A significant interface and lobe interaction was also observed for Alpha, $F(1.650, 244.221) = 9.186$, $p < .000$, $\eta_p^2 = .058$, Beta, $F(1.657, 245.197) = 9.613$, $p < .000$, $\eta^2 = .061$, and Theta, $F(1.637, 242.247) = 8.690$, $p = .001$, $\eta_p^2 = .055$ activity. This indicated that the differences in brain activity found between interfaces were not similar across all the lobes. For all frequency bands, the differences were greater in the Frontal and Parietal lobes compared to the Occipital lobe (see Figure 3-11 Average change in EEG brain activity from baseline by interface type and lobe).

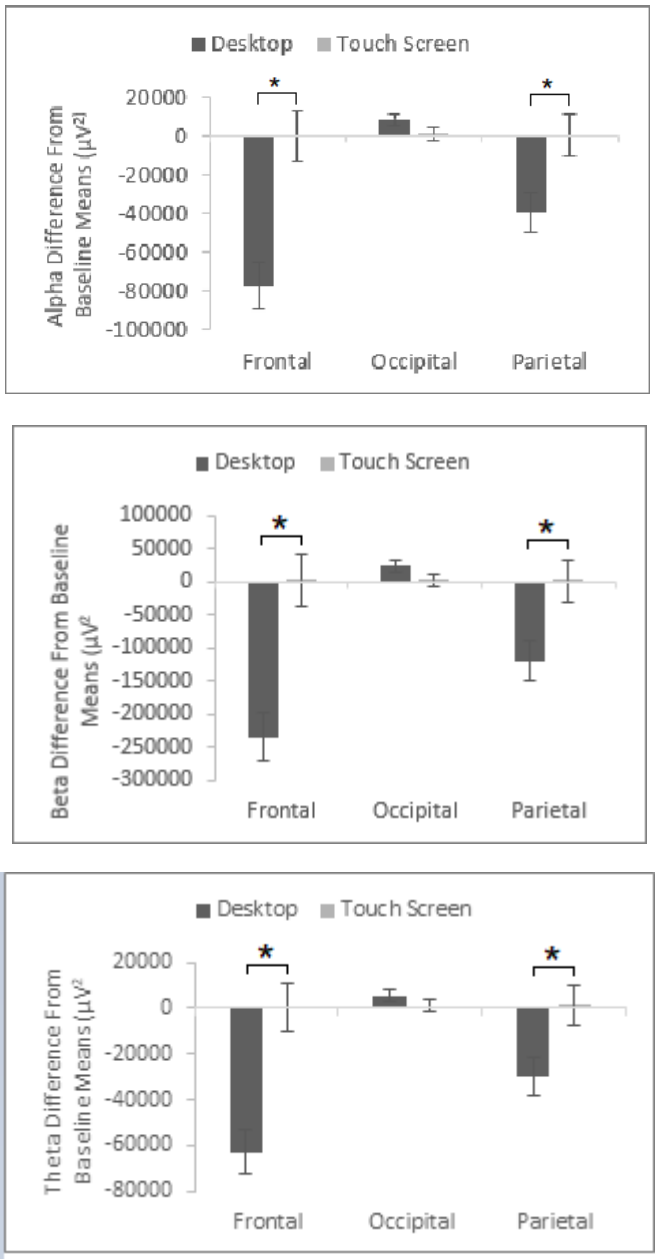
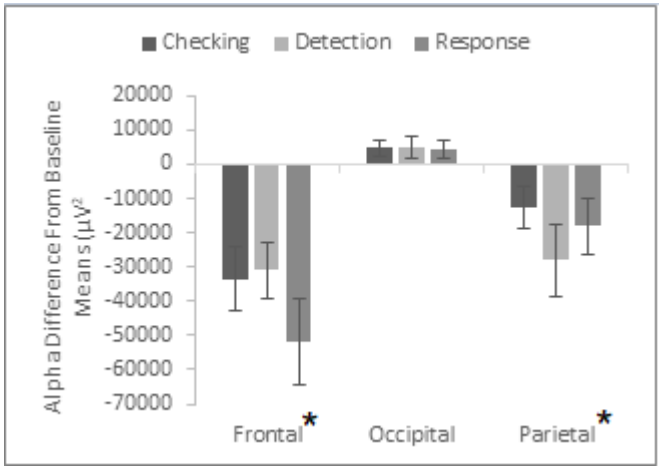
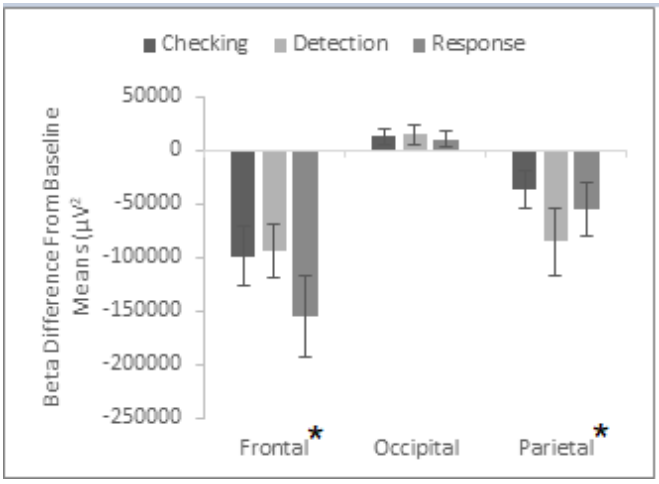


Figure 3-11 Average change in EEG brain activity from baseline by interface type and lobe (error bars denote standard errors, asterisks denote significant findings)

There was a significant interaction effect between task type and lobe for Alpha, $F(3.067, 453.908) = 3.351$, $p = .018$, $\eta_p^2 = .022$, Beta, $F(3.061, 453.014) = 3.363$, $p = .018$, $\eta_p^2 = .022$, and Theta, $F(3.084, 456.443) = 3.354$, $p = .018$, $\eta_p^2 = .022$ activity. This suggested that the magnitude of the difference in brain activity across the different tasks was not consistent for all lobes. Closer examination showed that the largest differences between task types were seen in the Frontal and Parietal lobes (see Figure 3-12 Average change in EEG brain activity from baseline by task type and lobe).





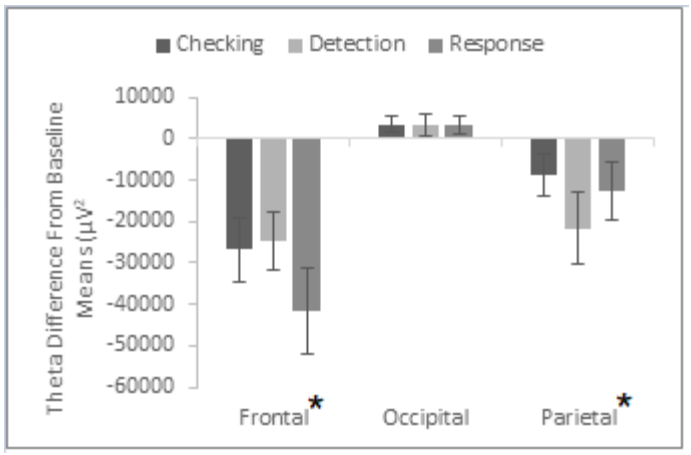
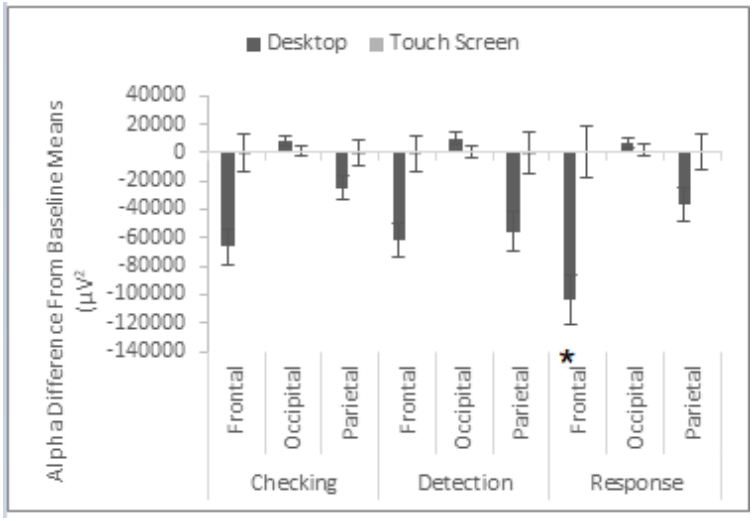


Figure 3-12 Average change in EEG brain activity from baseline by task type and lobe (error bars denote standard errors, asterisks denote significant findings)

Lastly, there was a significant interaction of task type, lobe, and interface type for Alpha, $F(3.067, 453.908) = 3.318$, $p = .019$, $\eta_p^2 = .022$, Beta, $F(3.061, 453.014) = 3.361$, $p = .018$, $\eta_p^2 = .022$, and Theta, $F(3.084, 456.443) = 3.153$, $p = .024$, $\eta_p^2 = .021$ activity. Further inspection of the results revealed that for all the frequency bands, the largest differences between interface types were seen in the Frontal lobe activity during the response implementation task. No interaction involving hemisphere or sites were significant (see Figure 3-13 average change in EEG activity by task, interface type and lobe, referenced to baseline).



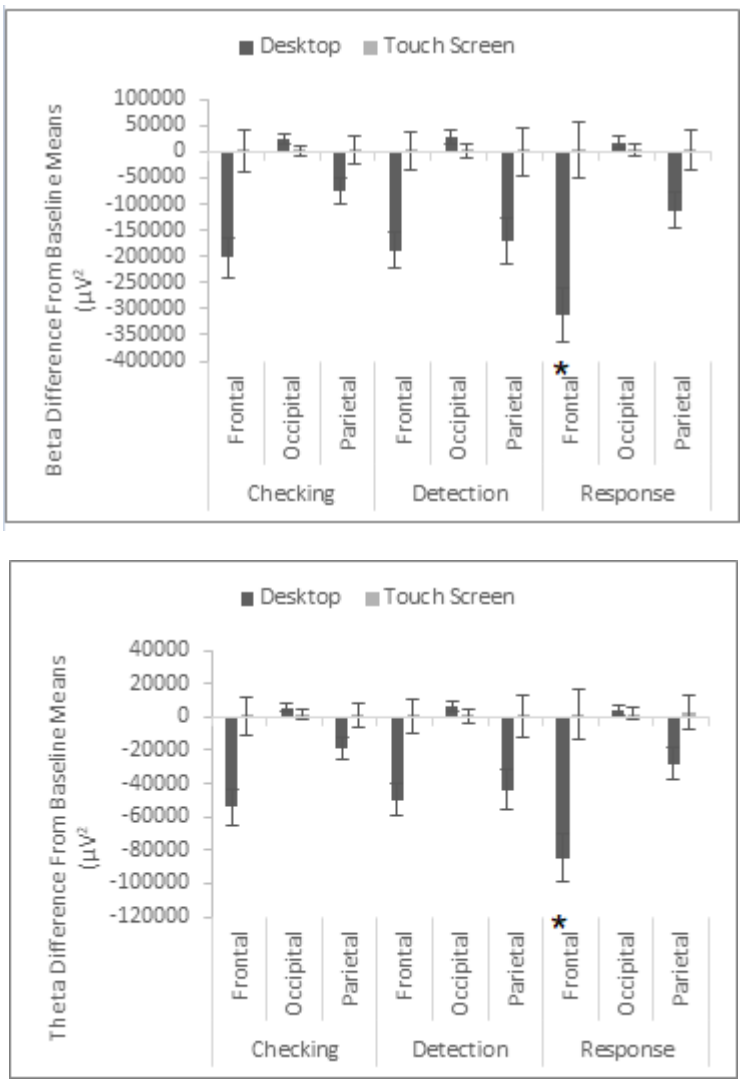


Figure 3-13 Avg. change in EEG brain activity from baseline by task, interface type & lobe (error bars denote standard errors, asterisks denote significant findings)

3.4.2.2.2 Transcranial Doppler (TCD)

A 3 (task type: checking, detection, and response implementation) × 2 (hemisphere: left and right) × 2 (interface type: desktop and touchscreen) mixed ANOVA was conducted to determine if there was an overall effect of task type and interface type on CBFV. The analyses also examined if the inter-hemisphere differences in CBFV were significantly different among the task types, and between interface types. Task type and lobe were repeated-measures variables and interface type was the between-subject variable.

A significant main effect was found for task type, $F(2, 288) = 5.602$, $p = .004$, $\eta_p^2 = .037$, such that the checking task ($M = .662$) resulted in significantly greater CBFV differences from baseline compared to the response implementation task ($M = -0.122$). In addition, a significant main effect for hemisphere was found, $F(1,144) = 4.881$, $p = .029$, $\eta_p^2 = .033$, such that the Left hemisphere ($M = 0.598$) resulted in significantly greater CBFV differences from baseline

compared to the Right hemisphere (*M* = -0.167). No other main or interaction effects were observed.

3.4.2.2.3 Functional Near-Infrared Spectroscopy (fNIRS)

A 3 (checking, detection, and response implementation) × 2 (left and right hemisphere) × 2 (interface type: desktop and touchscreen) mixed ANOVA was run to determine whether oxygenation was significantly different across task types and between interface types. The ANOVA interactions were examined to determine whether differences in oxygenation across tasks were the same for the two interface groups, and for the two hemispheres.

A significant main effect was found for task type, $F(1.764, 252.276) = 49.096$, $p < .000$, $\eta_p^2 = .256$, such that the detection task ($M = 1.093$) resulted in a greater increase in blood oxygenation from baseline compared to the checking ($M = 0.143$) and response implementation ($M = -0.101$) tasks. A significant main effect was found for interface type, $F(1, 143) = 57.721$, $p < .000$, $\eta_p^2 = .288$, such that the touchscreen interface group ($M = 1.477$) showed a greater increase in blood oxygenation from baseline compared to the desktop interface group ($M = -0.721$).

There was a significant main effect for hemisphere, $F(1, 143) = 11.262$, $p = .001$, $\eta_p^2 = .073$, such that the rise in oxygenation from baseline was overall higher in the Right hemisphere ($M = 0.567$) compared to the Left hemisphere ($M = 0.190$), and this was true across the different interface and task types. In addition, a significant interaction between task type and interface type was found, $F(1.764, 252.276) = 6.533$, $p = .003$, $\eta_p^2 = .044$, such that the Touchscreen interface group showed an average increase in blood oxygenation from baseline, whereas the Desktop interface group had an average decrease in blood oxygenation from baseline, and the differences between the interface groups were most pronounced in the Detection task (Figure 3-14 fNIRS difference in blood oxygenation as a function of task and interface type, referenced to baseline activity).

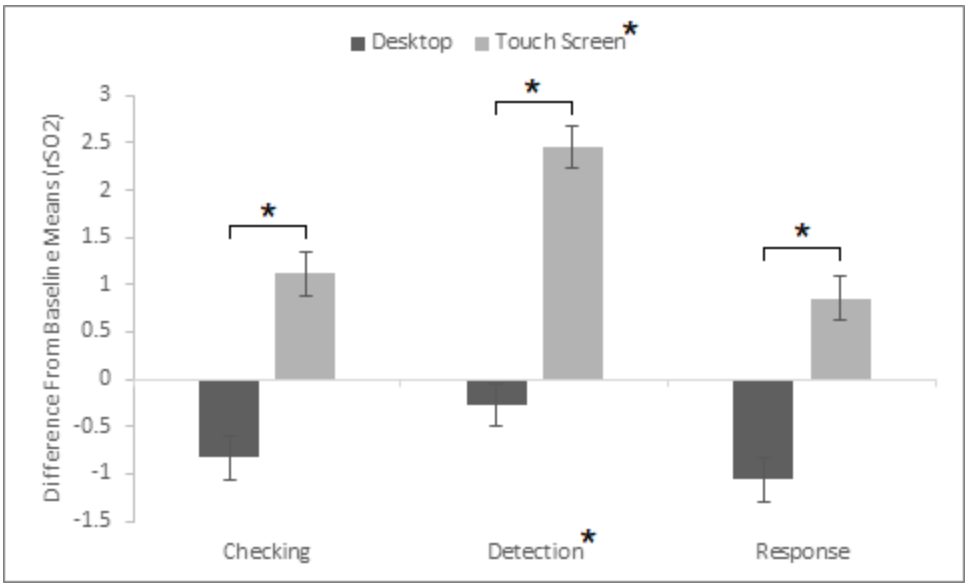


Figure 3-14 fNIR difference from baseline means by task type and interface type (error bars denote standard errors, asterisks denote significant findings)

3.4.2.3 Electrocardiogram (ECG)

A series of 3 (task type: checking, detection, and response implementation) × 2 (interface type: desktop and touchscreen) mixed ANOVAs was conducted to determine if the different task types and interfaces affected HR, HRV, and IBI. These analyses also assessed the interactive effects between the task types and interface types which would reveal if any differences found across task types were similar for the Desktop and Touchscreen groups. Task type was a repeated-measures variable and interface type was a between-subjects variable.

For HR, a significant main effect was found for task type, $F(1.808, 256.735) = 7.585$, $p = .001$, ${}_p\eta^2 = .051$, such that the checking task ($M = 1.290$) resulted in significantly greater increases in HR from baseline compared to the response implementation ($M = -1.850$) task type. A significant main effect was found for interface type, $F(1,142) = 8.833$, $p = .003$, ${}_p\eta^2 = .059$, such that participants using the desktop interface ($M = 1.846$) showed greater increases in HR from baseline compared to participants that used the touchscreen interface ($M = -2.378$).

For HRV, a significant main effect was found for task type, $F(1.789, 254.087) = 13.793$, $p < .000$, $\eta_p^2 = .089$, such that the response implementation task ($M = 12.205$) resulted in significantly greater increases from baseline compared to the checking ($M = 5.457$) and detection ($M = 3.651$) task types. A significant main effect was also found for interface type, $F(1,142) = 14.550$, $p < .000$, $\eta_p^2 = .093$, such that participants that used the touchscreen interface ($M = 13.370$) displayed greater increases in HRV from baseline compared to participants that used the desktop interface ($M = 0.838$). There was a significant interaction effect between task type and interface type, $F(1.789, 254.087) = 12.484$, $p < .000$, $\eta_p^2 = .081$, in which the largest differences between the interface groups were found for the checking and response implementation task (Figure 3-15 ECG HRV difference from baseline means by task type and interface type).

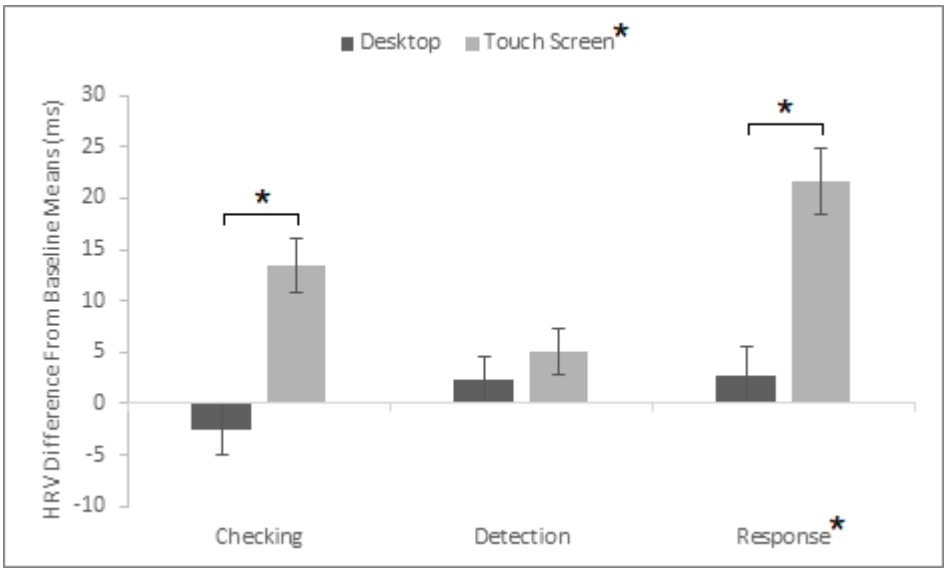


Figure 3-15 ECG HRV difference from baseline means by task type and interface type (error bars denote standard errors, asterisks denote significant findings)

For IBI, a significant main effect was found for task type, $F(2, 284) = 3.642$, $p = .027$, $\eta_p^2 = .025$, such that the checking task ($M = -28.959$) resulted in significantly greater decreases in IBI from

baseline compared to the response implementation (*M* = -21.390) task. Also, a significant interaction effect between task type and interface type was found, $F(2, 284) = 8.672$, $p < .000$, $\eta_p^2 = .058$, in which the largest differences between the two interface groups were observed during the checking and response implementation task (Figure 3-16 ECG IBI difference from baseline means by task type and interface type) (see Appendix C for a summary of results of physiological measures).

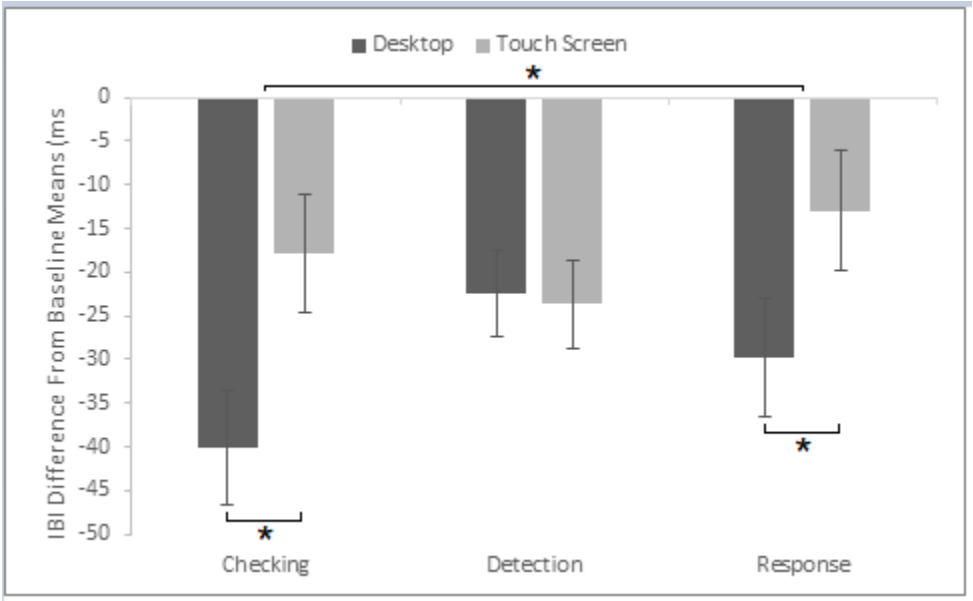


Figure 3-16 ECG IBI difference from baseline means by task type and interface type (error bars denote standard errors, asterisks denote significant findings)

3.4.3 Performance Measures

3.4.3.1 Communication Reporting

Communication reporting variables included percent correct, location help, clarification, and requests for repeating an instruction. Four 3 (task type: checking, detection, and response implementation) × 2 (interface type: desktop and touchscreen) mixed ANOVAs were conducted for each of the four measures to determine if there was a significant difference between task types and between interface types. The potential interaction of task and interface type on each of the four measures was also assessed. Task type was a repeated-measures variable and interface type was a between-subjects variable.

For percent correct, a significant main effect was found for task type, $F(2, 296) = 12.119$, $p < .000$, $\eta_p^2 = .076$, such that participants' performance was significantly better during the response implementation task (*M* = 90.985) compared to the checking (*M* = 86.752) and detection (*M* = 83.270) task types. A significant interaction between task type and interface was found, $F(2, 296) = 5.713$, $p = .004$, $\eta_p^2 = .037$, such that participants using the desktop interface performed better during the checking and response implementation task types, but not the detection task type (Figure 3-17 Percent correct means by task type and interface type). No significant effects were found for location help.

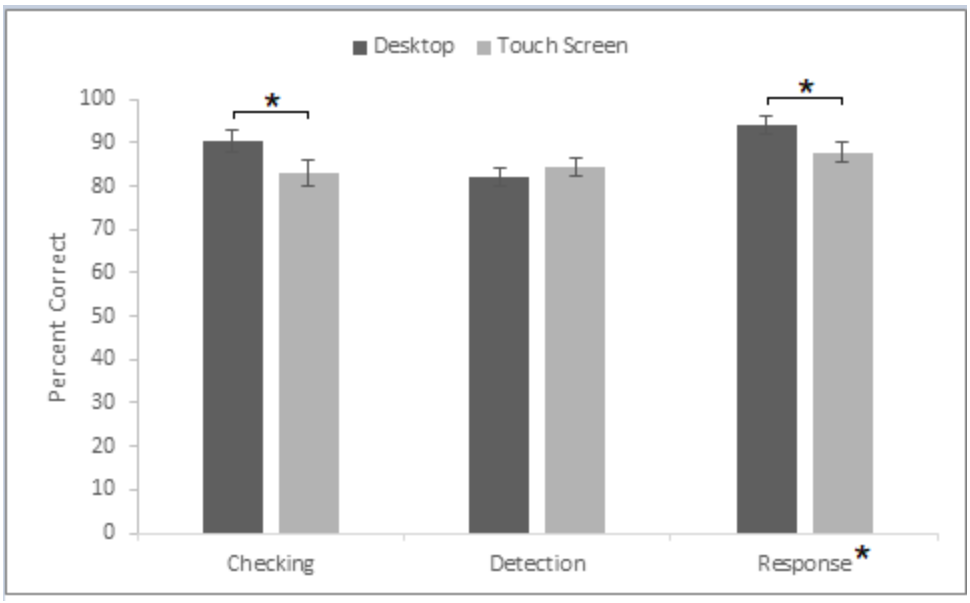


Figure 3-17 Percent correct means by task type and interface type (error bars denote standard errors, asterisks denote significant findings)

There was a significant main effect for task type on the number of clarifications, $F(1.507, 287.092) = 104.660$, $p < .000$, $\eta_p^2 = .414$, such that significantly more clarifications were needed during the detection task ($M = 1.988$) compared to the checking ($M = 0.472$) and response implementation ($M = 0.477$) task types.

For the number of repeats, a significant main effect was found for task type, $F(1.417, 209.728) = 91{,}700$, $p < .000$, $\eta_p^2 = .383$, such that significantly more repeats were needed during the detection task ($M = 1.354$) compared to the checking ($M = 0.258$) and response implementation ($M = 0.276$) task types.

3.4.3.2 Navigation and Identification

Navigation variables included: locating a correct control, the number of additional attempts to locate a correct control and locating a correct control on the first attempt. Three 3 (task type: checking, detection, and response implementation) × 2 (interface type: desktop and touchscreen) mixed ANOVAs were conducted for each of the three measures to determine if there was a significant difference between task types and between interface types. The analyses also revealed if the two interface groups showed similar patterns of differences in performance across the tasks. Task type was a repeated-measures variable and interface type was a between-subjects variable.

For locating a correct control, a significant main effect was found for task type, $F(1.790, 264.909) = 12.795$, $p < .000$, $\eta_p^2 = .080$, such that participants were able to correctly locate more controls for the response implementation task ($M = 3.721$) compared to the checking ($M = 3.126$) and detection ($M = 3.123$) task types. A significant main effect was found for interface type, $F(1, 148) = 6.249$, $p = .014$, $\eta_p^2 = .041$, such that participants were able to correctly locate more controls using the desktop interface ($M = 3.471$) compared to the touchscreen interface ($M = 3.176$). A significant interaction effect was found between task type and interface type, $F(1.790, 264.909) = 4.509$, $p = .012$, $\eta_p^2 = .030$, such that participants using the Desktop

interface were able to correctly locate more controls during the detection and response implementation tasks, but not in the checking task (Figure 3-18 Correct control means by task type and interface type).

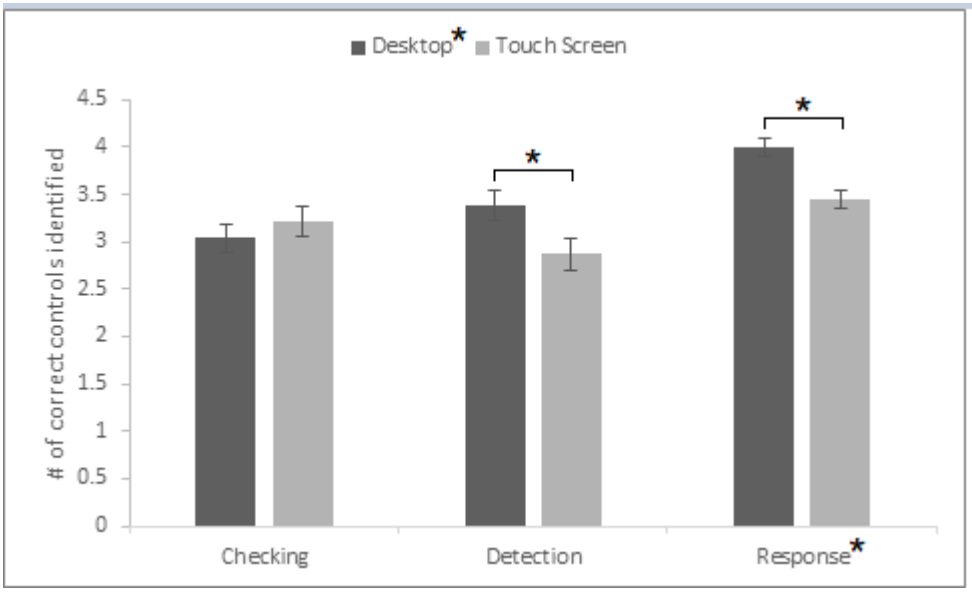


Figure 3-18 Correct control means by task type and interface type (error bars denote standard errors, asterisks denote significant findings)

For the additional attempts to locate a correct control, a significant main effect was found for task type, $F(1.004, 148.522) = 8.944$, $p < .000$, $\eta_p^2 = .057$, such that the detection task ($M = 6.659$) required significantly more attempts to locate a correct control compared to the checking ($M = 0.227$) and response implementation ($M = 0.542$) task types. A significant main effect was found for interface type, $F(1, 148) = 4.223$, $p = .042$, $\eta_p^2 = .028$, such that participants using the touchscreen interface ($M = 3.914$) required significantly more attempts to locate a correct control compared to participants using the desktop interface ($M = 1.037$). A significant interaction effect was found between task type and interface type, $F(1.004, 148.522) = 4.140$, $p = .044$, $\eta_p^2 = .027$, such that the differences between the touchscreen and desktop groups were most prominent during the detection task (Figure 3-19 Additional attempt means by task type and interface type).

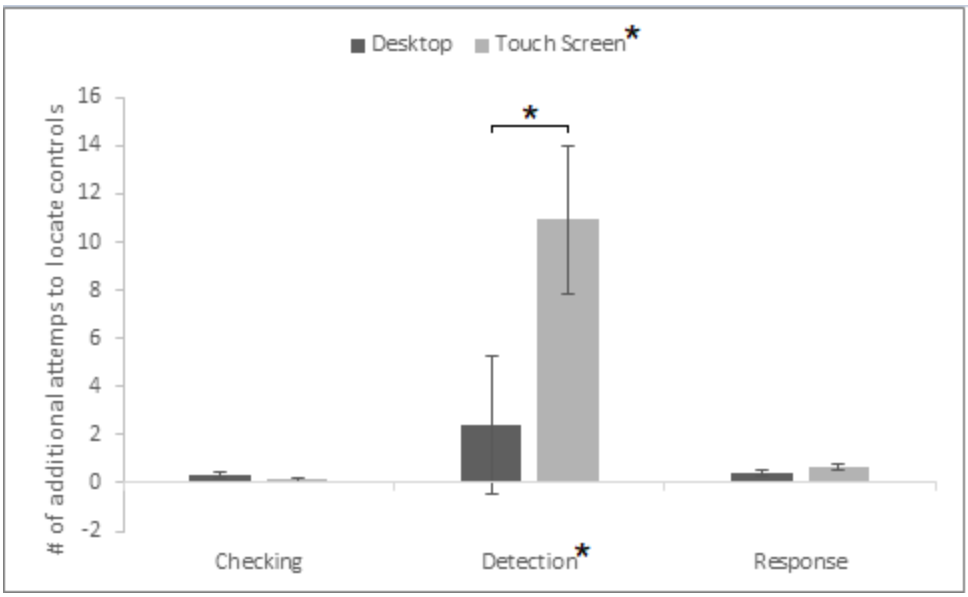


Figure 3-19 Additional attempt means by task type and interface type (error bars denote standard errors, asterisks denote significant findings)

For locating a correct control on the first attempt, a significant main effect was found for task type, $F(1.848, 273.477) = 49.391$, $p < .000$, $\eta_p^2 = .250$, such that, in general, participants had significantly more difficulty finding the correct control on the first attempt during the detection task ($M = 0.173$) compared to the checking ($M = 0.575$) and response implementation ($M = 0.605$) tasks. A significant interaction effect was found between task type and interface type, $F(1.848, 273.477) = 11.191$, $p < .000$, $\eta_p^2 = .070$, such that participants using the desktop interface were able to correctly locate more controls on the first attempt during the checking and detection tasks, but not in the response implementation task (Figure 3-20 Percent of controls located on the first attempt means by task type and interface type).

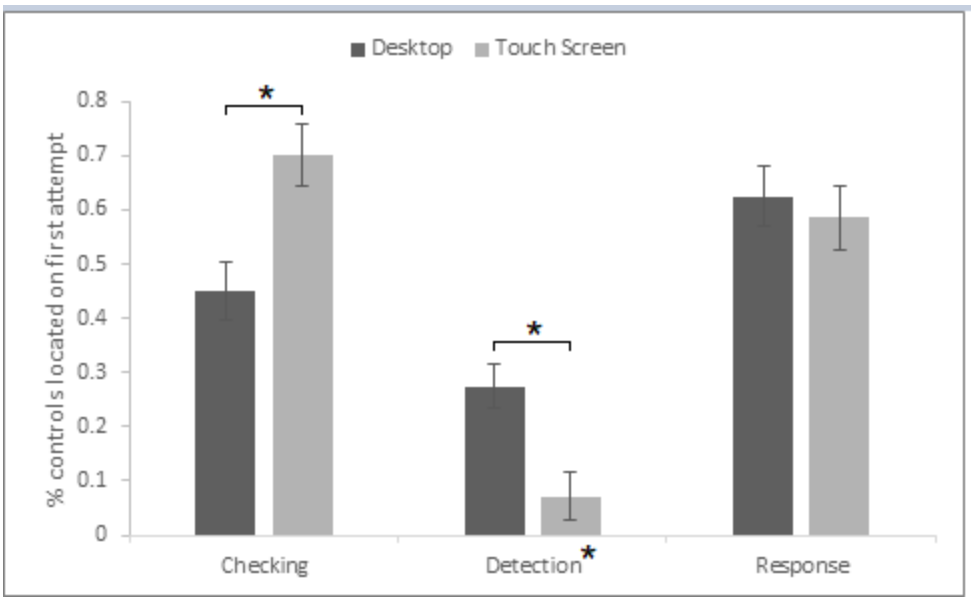


Figure 3-20 Percent of controls located on the first attempt means by task type and interface type (error bars denote standard errors, asterisks denote significant findings)

3.4.3.3 Action

Independent sample *t*-tests were conducted to determine if there were significant differences between interface types for various action performance variables. Below are the descriptions and results of each action performance measure for the detection and response implementation task types.

3.4.3.3.1 Detection

The percentages of correct responses, missed events, and false alarms for each participant were measured while completing the detection task. A significant effect for interface type was found for the percent of correct responses, $t(148) = 4.659$, $p < .000$, such that participants using the desktop interface ($M = 64.816$) responded with significantly more percent correct responses compared to participants using the touchscreen interface ($M = 47.946$). A significant effect for interface type was found for the percent of missed events, $t(148) = -4.585$, $p < .000$, such that participants using the desktop interface ($M = 27.522$) responded with significantly less percent missed events compared to participants using the touchscreen interface ($M = 43.023$). No significant effect was found for the number of false alarms.

3.4.3.3.2 Response Implementation

The percentages of correct responses, description error, mode error, and number of mis-order errors for each participant were measured while completing the Response Implementation task. A significant effect for interface type was found for the percent of correct responses, $t(148) = 5.961$, $p < .000$, such that participants using the desktop interface ($M = 60.298$) responded with significantly more percent correct responses compared to participants using the touchscreen interface ($M = 27.650$). A significant effect for interface type was found for the percent description error, $t(148) = 2.809$, $p = .006$, such that participants using the desktop interface (*M*

= 2.060) responded with significantly more percent description error compared to participants using the Touchscreen interface (M = 0.204). A significant effect for interface type was found for the percent mode error, $t(148) = -10.428$, $p < .000$, such that participants using the desktop interface (M = 4.834) responded with a significantly lower percent mode error compared to participants using the touchscreen interface (M = 31.314). A significant effect for interface type was found for the number of mis-order errors, $t(148) = -5.145$, $p < .000$, such that participants using the desktop interface (M = 2.763) responded with significantly fewer mis-order errors compared to participants using the touchscreen interface (M = 5.429) (see Appendix D for a summary of results of performance measures).

3.5 Discussion

The present study set out to determine the feasibility of investigating issues of training, workload, and performance for operations within an NPP MCR, using novice participants, and a light weight but realistic NPP simulation approach. The present study also examined how different interface types impacted subjective and physiological workload, and performance.

Three common control room tasks (checking, detection, response implementation) were operationally defined and used for the performance-based assessment. Each task type consisted of four steps that were executed using three-way communication led by the experimenter acting as the SRO. A display type manipulation was introduced as an independent variable to compare the performance outcomes and workload associated with touchscreens and point and click mouse and desktop interaction style. Future control rooms could utilize either interaction design, thus, understanding the potential limitations of each is important for continued assurance of safe operations in such integrated control rooms.

Overall, the results suggested that numerous subjective and physiological measures were sensitive to detecting workload changes between the three task types. Generally, compared to all others, NASA-TLX frustration and fNIRS were more sensitive indices.

With respect to the interface variable, the results suggest several potential human factors challenges for touchscreen technology (depending on the implementation), such as more required body and figure movement, manipulation difficulty for I&Cs, and difference in physical feedback. Specifically, the touchscreen interface required participants to move around and use their fingers to directly manipulate controls (the desktop interface relied on a translated input through a mouse). Such frequent body and figure movement may lead to arm fatigue and obstruction by finger (or fingerprints). The virtual buttons on the touchscreen interface were more difficult to manipulate, especially when they occurred below the shoulder level of the participant, a result supported by the higher rating on the NASA-TLX subscale for performance. However, the results of the NASA-TLX suggest that the touchscreen interface may have been perceived as easier to use (lower rating on the effort subscale). Similarly, the reduced reported demand on short-term memory and selected spatial processing for the Touchscreen interface type suggests a perceived ease of use, separate from observed performance.

Recent research suggests that comparing to a sitting position, standing position may lead to more misses and longer reaction time when interacting with buttons in smaller size (Chourasia et al., 2013); pilots' physical effort varies depending on the location of touch screen and task duration (Barbé et al., 2012). Future research is needed to understand these kinds of ergonomics issues in the NPP domain. Unlike the desktop interface (mouse cursor can be positioned correctly over the button prior to clicking it), participants using the touchscreen

interface can only get feedback about the accuracy of their touch after they had performed the action. Regardless of interface types, the detection task type elicited the highest workload of all three tasks.

Study 1 provided a systematic baseline experiment for understanding workload as it occurs in an NPP MCR and determined the measures suitable for detecting workload responses. The experiment also provided evidence that the measures were sensitive to the manipulations with novice participants. Novice participants were trained on performing the basic tasks in an NPP simulator.

3.5.1 Workload

The next several sections will highlight key findings from the subjective and physiological measures of workload. Each type of workload assessment provides a unique assessment of different dimensions and timescales of workload.

3.5.1.1 Subjective Workload Measures

3.5.1.1.1 NASA-TLX

Participants rated their global workload consistently higher for detection compared to checking and response implementation tasks. This higher rating can be attributed to the higher ratings in the frustration dimension, reflecting higher frustration with performing detection, relative to the other tasks. The detection task was like a vigilance task in that the participants had to sustain attention for a prolonged interval (i.e., participants had to monitor a gauge for changes in the reading for five minutes and click on the “acknowledge” button whenever they detected a change in level), which is known to increase workload. While the results are consistent with a vigilance task, this was not a vigilance decrement task. Rather, the experimental task was operationalized from a real-world NPP task and there were not sufficient trials to test for vigilance, as there were only 60 changes that occurred randomly throughout the five minutes. As the participants performed all four detection tasks in succession, they essentially monitored four different gauges continuously for twenty minutes and had to detect 240 level changes in the gauges. In contrast, the checking and response implementation tasks did not involve vigilance and were much shorter in duration, and so did not elicit as much frustration. The high frustration experienced during the detection task concurred with similar past findings of elevated frustration during vigilance tasks (Szalma, 2014; Warm et al., 1996).

The desktop interface elicited higher workload overall compared to the touchscreen interface, and the differences between interface groups were most marked for Effort and Performance workload. Workload due to Effort was lower while perceived Performance was higher in the touchscreen group relative to the desktop group. The touchscreen interface did not require as much effort to use, and those who used it tended to rate their performance higher than those who used the desktop interface. It should be noted that these findings are slightly contradictory to the actual performance differences found between the two interface types which will be discussed in more detail following the workload results.

3.5.1.1.2 ISA

Unlike the results from the NASA-TLX, the workload assessed by the ISA showed that the highest level of workload was experienced in the checking task. This could be because the

aspect of workload tapped by the ISA pertained more to the pace of the task (i.e., participants were instructed that a rating of 2 represented that they felt relaxed and that they had more than enough time for the task, and a rating of 3 represented that they felt like the task progressed at a comfortable busy pace). This corroborates with the higher mean rating on the NASA-TLX's Temporal Demand during the Checking task compared to the other two tasks.

3.5.1.1.3 MRQ

Nine of the 14 items that are relevant to the present study showed workload differences among the tasks. The spatial attentive, spatial positional, visual lexical, and vocal processes were most highly activated by the checking task, while the auditory emotional, visual phonetic, and vocal processes were most highly activated by the response implementation task. On the other hand, the spatial attentive, spatial concentrative, spatial quantitative, and visual temporal processes were most highly activated by the detection task compared to the other tasks.

The results between the two interface groups generally mirror one another and were in line with expectations given the nature of the three tasks. For instance, the finding that the checking task elicited the most workload related to the spatial positional process probably reflects the extent to which determining precise spatial locations featured in the checking task. While all tasks required participants to determine precise spatial locations, the checking task solely consists of spatial positional processing. Spatial positional processing comprised only a part of the detection and response implementation tasks. Hence, the demand for spatial positional processing is more salient for the checking task compared to the other two tasks.

Spatial attentive processing was lowest in the response implementation task probably because of the location of the controls in each task type. For the checking and detection tasks, participants were instructed to locate controls that were spread across both panels. However, for the response implementation task, participants were instructed to locate controls that were within the same panel and in relatively close in proximity to each other. As a result, participants were required to use less spatial attentive processing during the response implementation task because they were required to locate controls over a smaller area of space. This finding concurs with the navigation performance results which will be discussed in section 3.6.1 (p. 84).

Visual temporal processing was highest in the detection task compared to the checking and response implementation task types because the detection task required participants to interact with gauges by identifying small changes that occurred every few seconds. Overall, participants were required to judge 60 changes during a 5-minute interval for each of the gauges. This resulted in 240 changes occurring across the entire task. The checking and response implementation task types did not require judging the timing of events as participants were only required to interact with valves by reporting their state or opening/closing a switch respectively.

Spatial quantitative processing was also highest in the detection task compared to the checking and response implementation task types because the detection task required a high amount of spatial quantitative processing to determine each gauge level, while the checking and response implementation task did not require identification of a numerical quantity. Spatial concentrative processing was highest in the detection task because while all tasks required a similar amount of spatial concentrative processing during the navigation component, the detection task required additional spatial concentrative processing once a control was located. The detection task required participants to identify when a non-digital gauge reached a particular level. To complete this task, participants had to determine the numerical value of each dash by identifying

the increments of the spaced dashes between gauge values. For example, one gauge contained the numbers 0, 500, 1000, 1500, 2000, 2500, and 3000, each with nine dashes in between. In this case, each dash was an increment of fifty. On the other hand, another gauge contained the numbers 0, 100, 200, 300, 400, 500, 600, and 700, each with four dashes in between. In this case, each dash was an increment of twenty. The differences in gauge design, coupled with the fact that the spacing between the dashes varied per gauge, attributed to a high amount of spatial concentrative processing during the detection task.

Vocal processing was highest for the response implementation and checking tasks compared to the detection task because, while all three task types required a similar amount of voice usage via three-way communication, the checking and response implementation tasks were executed more quickly and had shorter breaks between the communication parts. During the checking and response implementation tasks, participants were communicating with the SRO at a faster pace compared to the detection task. The detection task entailed monitoring a gauge's state without requiring communication with the SRO for a five-minute period.

When the ratings for each MRQ item were examined in the context of the effects of the interfaces, the results showed that desktop interface resulted in greater activation of short-term memory, spatial attentive, and spatial categorical processes. This is likely due to the limitations of the amount of information that could be simultaneously displayed on the desktop UI, participants using the desktop interface required to remember the locations of the controls as they scroll, pan, and zoom. Furthermore, the interaction effect between task type and interface type showed that spatial positional processing was highest for participants that used the desktop interface during the detection task compared to participants that used the touchscreen interface during the detection task. This finding is consistent with the spatial categorical and spatial attentive subscales results. Participants who used the touchscreen interface had less trouble determining the locations of the controls because they did not have to use the scroll, pan, and zoom functionality.

Taken as a whole, the findings indicate that the detection task with the desktop interface had the highest resource demands, as assessed by the MRQ.

3.5.2 Physiological Workload Measures

3.5.2.1 EEG

Across interface types, brain activity as measured by EEG showed a general increase in the occipital lobe and decrease in the frontal and parietal lobes from baseline. There were no differences in the level of brain activity across the three tasks. It was possible that the EEG measure was not sensitive to the differences among the three tasks, all of which shared common components such as locating the controls.

However, across all three tasks, the touchscreen interface group showed an increase in overall brain activity from baseline levels, whereas the desktop group showed a decrease in brain activity from baseline levels. This was true for alpha, beta, and theta frequency bands. Closer examination revealed that this difference between interfaces was most distinct in the frontal and parietal lobes during the response implementation task. The response implementation task required participants to perform gestures (i.e., click or touch, then hold and rotate) to open or shut specific valves. There are two potential explanations for this finding, first, the increased brain activation in the touchscreen group during the response implementation task could be

caused by the touchscreen interface requiring participants to move around to locate the controls and involved them directly manipulating controls with their finger instead of sitting at a desk and relying on a translated input such as a mouse. The response implementation task was also the task that required the most “elaborate” gesture, involving touching and holding down the valve’s “handle” and rotating it at least 45 degrees with a drag gesture to the left or right (left to shut, right to open), then releasing the handle. The differences among the tasks were most pronounced when comparing them by interface type. The second potential explanation is that an increase in alpha band power, in some circumstances can indicate a reduction in workload, however, this reduced alpha is typically observed in the context of a flat or elevated theta power (Borghini, Vecchiato, Toppi, Astolfi, Maglione, Isabella, Caltagirone, Kong, Wei, Zhou, Polidori, Vitiello, & Babiloni, 2012). Ball, DellaNoce, Quek, and North (2006) suggest that even if the gestures used for a touchscreen interaction are complex, it still may elicit a lower workload because those gestures are intuitive and do not require the translation from action in a horizontal plane (mousing) to interaction with digital objects in a vertical plane. Given the complexity in the spectral power indications of workload and the varied display types, further research would be needed to uncover the drivers of the changes in workload as a function of the display and interaction type changes.

EEG is a well-established, but fairly intrusive means of measuring the neural correlates of cognitive activity and task performance. However, under real world circumstances, where tasks are complex and share some similarities in terms of the behavioral execution of task performance, EEG may not be the most diagnostic measure. The diagnosticity of EEG is to some extent determined by study parameters and therefore might not be the best option for examining human factors impacts, such as changes in operator workload in a control room environment.

3.5.2.2 TCD

The TCD results were similar across the interface types. The change in CBFV from baseline levels was greatest during the checking task compared to the other tasks. In addition, the change in CBFV from baseline was generally greater in the left hemisphere relative to the right. Nevertheless, these findings do not lend themselves to easy interpretation as there was large variability observed in the CBFV values.

3.5.2.3 fNIRS

Relative to the other tasks, the detection task elicited the greatest change in regional oxygen saturation, rSO_2, from baseline levels. Previous research has shown that rSO_2 values increase with mental effort (Reinerman-Jones et al., 2014) as well as time on task for vigilance tasks (Funke et al., 2010). These results corroborate the findings from the subjective measures in that the detection task was found to elicit the highest workload ratings due to the strong vigilance component in the detection task. Furthermore, the rSO_2 change from baseline in the right hemisphere was greater than that in the left. This right hemispheric lateralization has been shown in vigilance tasks (e.g., Helton et al., 2010). Results of the differences in rSO_2 between the interface groups were similar to that for brain activity; the touchscreen interface group showed increased rSO_2 from the baseline while the rSO_2 in the desktop interface group decreased from baseline levels.

3.5.2.4 ECG

Compared to the other tasks, the checking task elicited greater changes in HR and smaller changes in HRV and IBI from baseline. Previous research has shown that increased HR, decreased HRV, and decreased IBI are linked to increases in workload (Wilson, 2002; Veltman & Gaillard, 1996). These findings corroborate the NASA-TLX temporal demand and ISA findings in that the higher level of workload induced by the checking task related to the quicker pace of the checking task compared to the other tasks.

The touchscreen interface group seemed to experience lower workload as indicated by their lower HR and higher HRV change from baseline scores relative to the desktop interface group. This difference was most apparent during the checking and response implementation tasks which showed that the touchscreen group had higher HRV and IBI. These results are in line with the NASA-TLX global workload score, various MRQ measures, and the EEG results and may be an indication that participants found the touchscreen easy to use because pointing is an intuitive and naturalistic gesture and the touchscreen did not require translation from horizontal to vertical planes (see Borghini, Astolfi, Vecchiato, Mattia, & Babiloni, 2012; Ball, DellaNoce, Quek, & North 2006). However, ease of use does not necessarily indicate better or more accurate performance, simply that the demands on the participants were lower in the touchscreen condition relative to the desktop condition.

3.6 Performance

3.6.1 Navigation

Measures of navigational ease included (i) the number of controls correctly located, (ii) number of additional attempts in locating controls, (iii) percent of controls correctly located on the first attempt.

The differences in navigation performance found among the three tasks were most likely due to the location of controls involved with the different tasks. For the detection task, participants were required to locate half of the controls on the A2 panel and half of the controls on the C1 panel. On the other hand, the checking task required participants to locate 75% of the controls on the A2 panel and 25% on the C1 panel. For the Response Implementation task, all of the controls were located in the A2 panel. Hence, as expected, the number of controls correctly located was highest for the response implementation task, which involved locating valves that were located within the same panel. This is consistent with the finding of lower reported demand on spatial attentive processes during the response implementation task. In contrast, navigation seemed most demanding during detection. The number of additional attempts to locate controls was higher, and the percent of controls correctly located on the first attempt was significantly lower during the detection task compared to the checking and response implementation task types.

Comparing the two interface types, navigation seemed easier with the desktop interface in general. The number of controls correctly located was higher and the number of additional attempts to locate controls was lower in the desktop group relative to the touchscreen group. This is despite the fact that there was no need for the touchscreen group to scroll-pan-zoom. It is possible that the visual complexity of viewing the entire panel, as well as the large size of the touchscreen display requiring movement to locate the controls made navigation more difficult for the touchscreen group. The increased ease in navigation with the desktop interface compared to the touchscreen interface as measured by the number of correctly located controls and the

number of additional attempts to locate control were most marked in the detection and response implementation tasks. However, the higher number of controls located correctly on the first attempt found with the desktop interface was most pronounced in the checking and detection tasks.

The finding of better navigational performance in the desktop interface group is decoupled from subjective and some physiological measures of workload. Participants using the touchscreen interface reported less workload and greater confidence about their performance, however, observed navigation performance favors the desktop UI. Nevertheless, the results of navigational performance seemed to correspond to that of past research which reported that desktop interfaces using a mouse interface result in better navigation performance and object identification performance compared to touchscreen interfaces (Sears & Shneiderman, 1991; Ulrich, Boring, & Lew, 2015).

3.6.2 Communication Reporting

Performance on the three-way communication reporting protocol was assessed by (i) percent correct communications, (ii) number of clarifications made, and (iii) number of times instructions were repeated. Percent correct communications was highest for the response implementation task and lowest during the detection task. These differences in communication performance were unlikely to be the outcome of any differences in communication reporting across the tasks as the language, format, and complexity of communication was virtually identical across the tasks. Instead, the results may reflect (a) the differences in the difficulty of locating the controls across the tasks, and (b) the differences in the duration of the tasks, both of which are associated with higher reported workload. The controls for the detection task were spread 50%-50% across the two panels, while, for the checking task, 75% of the controls were found in one panel and the other 25% of the controls were located in the other panel. On the other hand, all the controls for the response implementation task were located in the same panel. This reduction of the task load associated with locating controls may have enabled more resources to be allocated to communication reporting, resulting in better communication performance on the response implementation task.

Communication had a substantial influence on task duration. The number of clarifications required, and the number of times instructions were repeated seemed associated with the duration of each task. In the case of detection, this increase in task duration appears to be associated with perceived and observed task difficulty, corroborated with the notion that the lengthier duration of the detection task increased difficulty of the communications reporting. There were more clarifications and repetitions of instructions required during the detection task compared to the other tasks. In addition, this ease of communication found with the checking and response implementation tasks was more distinct from the desktop interface group.

Taken together, these results suggest that the workload burden imposed by the communication task needs to be contextualized with other task-related factors.

3.6.3 Performance on the Tasks

To better understand the performance differences between the interfaces, actual task performance was classified according to the type of controls the task was performed with. This is because the three tasks involve different numbers of gauges and valves.

3.6.3.1 Gauge Events

The desktop interface group showed higher percent correct responses and lower percent of misses with gauge events compared to the touchscreen group. As mentioned previously, these results are consistent with past research that demonstrated that touchscreen interfaces result in higher error rates for object identification performance compared to mouse interfaces (Sears & Shneiderman, 1991; Ulrich, Boring, & Lew, 2015).

3.6.3.2 Valve Events

Participants using the desktop interface had significantly higher correct responses and percent description errors, while participants using the touchscreen interface were found to have significantly higher percent of mode error as well as mis-order errors. Taken as a whole, these results are consistent with the navigation and gauge event results in that the desktop interface resulted in lower overall performance error compared to the touchscreen group. Although the desktop interface group showed a higher percent of description errors, this error rate was still low and can be attributed to the performance of a small minority (2%) of participants.

3.7 Conclusion

3.7.1 Overview of Study 1 Findings

Study 1 confirmed the feasibility of using novice participants to perform common NPP operator tasks (i.e., checking, detection, and response implementation) in a simplified desktop- or touchscreen-based simulated environment. The fact that only one participant was dismissed from the experiment due to failure to reach proficiency on the progressive training module, suggests that university students (i.e., novice participants) can gain proficiency in realistic (rule-based and skill-based) operator tasks in a simplified controlled environment. The detection task induced the greatest workload and resulted in the poorest performance, indicating the detection task required more decision-making and cognitive resources. The similar and overall low reported level of subjective workload for checking and response implementation tasks suggests that the fine motor response required to complete the response implementation task type does not add additional workload beyond that associated with the checking behavior. Numerous subjective and physiological measures were sensitive to workload changes across the three task types. NASA-TLX frustration and fNIRS were the most sensitive measures. However, fNIRS showed an opposite workload trend (supported by TCD, ECG), and some of the other physiological findings were mixed or inconclusive indicating physiological measures may not be reflective of task demand, but rather the task components themselves (i.e., the time required to complete each task) or limitation of those measurement techniques.

Overall, the traditional desktop and touchscreen interfaces showed similar trends in terms of both workload and task performance. The workload was highest in the detection task, which was longer in duration and required participants to be vigilant. Several subtle workload performance differences between checking and response implementation task types were found as indicated by the ISA, MRQ, ECG, and performance measures. Checking had higher temporal demand due to the rate in which the task was completed. Response implementation was able to be performed more effectively due to the location of the controls. The mapping of task and context attributes onto workload findings is a positive indication that the measures selected are appropriate for assessing the workload associated with common control room tasks. Additionally, understanding the fundamental cognitive and behavioral features underlying each

task is critical in selection of an appropriate workload measure, or a subscale within a particular measure.

Despite the observation of lower levels of workload generally, there were several human factors challenges associated with the use of the touchscreen interface. In the detection task, the touchscreen interface induced more brain activity and was related to more control location errors and a higher frequency of missed events. Regarding increased brain activity for touchscreen, more research would be needed to determine if these results are generalizable or due to limitations in the study design or other not yet quantifiable factors.

Regarding the control location errors, maximum control of skilled movements and speed of operation involving the hands and arms for jobs held in front of the body is achieved by holding the elbows down to the sides and the arms bent at right angles. This may explain the increase in location errors when working with touchscreen interfaces over traditional desktop interfaces, this finding is consistent with the touchscreen findings from Ulrich, Boring, and Lew, 2015.

The touchscreen interface required participants to move around and use their fingers to directly manipulate controls (the desktop interface relied on a translated input through a mouse). Although the touchscreen interface reduced the demand on short-term memory and selected spatial processing, the benefits of these load reductions were offset by the inaccurate manipulation of the touchscreen's buttons. Button manipulation issues were particularly pronounced when the buttons were located below the shoulder level of the participant, likely due to issues of vision or line of sight, biomechanics, or perhaps a combination. Unlike the desktop interface, participants using the touchscreen interface can only get feedback about the accuracy of their touch after they had performed the action. This could explain the performance difference, but it could result from mouse control allowing for finer movement and control in general. More research is needed to confirm the underlying cause for these results seen in the present work.

3.7.2 Conclusions for Study 1

Study 1 provided the baseline human performance data needed to establish the proof of concept.

3.7.2.1 Part 1 of Establishing the Proof of Concept

Broadly speaking, establishing the proof of concept was determining the following: Can novice participants perform operator tasks in a MCR environment? As a first step to answering this question, the research team:

- Created a cognitively similar environment to an NPP MCR with enough fidelity that the cognitive processes engaged by participants are comparable to those in real operators through a combination of applied experimental research and expert operator elicitation techniques.
- Demonstrated that novices can successfully perform realistic operator tasks within the proof-of-concept environment.

3.7.2.2 Baseline Human Performance Data

The purpose of the present research effort was to systematically collect human performance and workload data when performing critical tasks in NPP MCR while utilizing both a mouse-click

desktop and touchscreen interface. The results comparing the two interface types displayed similar trends in terms of subjective workload, objective workload, and performance associated with the three types of rule-based (Rasmussen, 1983) NPP MCR tasks.

3.7.2.2.1 If novices can perform proficiently, are there differences in performance as a function of task type and what did we learn about the workload associated with each of the task types?

The detection task was found to result in the highest levels of workload and lowest levels of performance compared to the checking and response implementation task types. Results demonstrated that the workload and performance differences between the two tasks were somewhat negligible. However, for the touchscreen interface, several subtle workload and performance differences were found between the checking and response implementation task as indicated by the ISA, MRQ, ECG, and several performance measures. Closer examination of the results highlighted differences in certain task characteristics. The checking task had higher temporal demand due to the rate in which the task was completed, and the response implementation task was able to be performed more effectively due to the location of the controls. These differences, which persisted across interface types, were probably a result of relatively small task effects that required a large sample size to be detected. Combining the datasets provided that large sample. Regardless, the research efforts provide a clear indication that despite the quicker pace of the checking task, and the added requirement of having to manipulate the valves in the response implementation task, the workload was still highest in the detection task which was longer in duration and required participants to sustain attention for a long period of time.

3.7.2.2.2 What did we learn about the impact of different interfaces?

Results on the effects of interface types yielded slightly contradictory results on the surface. Although some subjective measures (e.g., NASA-TLX and several MRQ items) indicated that the touchscreen interface was associated with lower overall workload and did not place as much demand on short term memory and certain spatial processes as the desktop interface did, the touchscreen interface induced more brain activity and was related to more control location errors and higher frequency of missed events in the detection task. This pattern of results highlights the importance of collecting both subjective and objective workload measures. The perceived workload associated with using the touchscreen interface was reportedly low, but the performance-based and physiological measures tell a slightly different story. It is likely that the touchscreen interface, in requiring participants to move around and use their fingers to directly manipulate controls instead of relying on a translated input through a mouse, involved the participants more than the desktop interface. By enabling all the controls to be visible, the touchscreen interface also reduced the demand for short-term memory and selected spatial processing. However, the buttons on the touchscreen interface were reported to be more difficult to reach, especially when they occurred below the shoulder level of the participant. In addition, unlike the desktop interface, participants using the touchscreen interface would only get feedback about the accuracy of their touch after they had performed the action. Participants using the desktop interface could ensure that their mouse cursor was positioned correctly over the button prior to clicking it.

Large touchscreens with entire I&C panels were found to be beneficial to increase physical fidelity and reduce response time, but also induced more missed touch errors when compared to performance using desktop interface with mouse click input. These results suggest that a

future research focus on optimizing touchscreen controls for improved performance may be beneficial.

3.7.2.2.3 What did we learn about assessing workload using novice participants in the nuclear domain

These results contribute both to practice and theory. From a practical standpoint, conducting human performance research on the tasks associated with a highly specialized job (e.g., pilots, ROs in an NPP MCR) has always been a challenge. However, the present research efforts have demonstrated that some of these challenges can be surmounted through a systematic research approach that includes (i) understanding the nature of the RO's job and the associated tasks, (ii) selecting tasks which can be "scaled" down appropriately to still maintain a level of fidelity to the actual tasks to ensure generalizability of results, but which can be performed by novice participants, and (iii) careful design of training protocol for the novice participants and experimental protocol that would address pertinent research questions. Future research with actual ROs can help to validate this approach. Furthermore, the research contributes to theoretical understanding of workload and performance assessment by demonstrating that workload is multifaceted and different measurement techniques will more effectively highlight different facets of the construct. From a practical standpoint, the availability of multiple workload measures enables practitioners' selection of a set of measurement tools that are well matched to the fundamental characteristics of the operator's task. This technique to facet matching maximizes sensitivity to variations on workload.

4 STUDY 2

The purpose of Study 2 was to validate the findings of Study 1 with a small group of former operators. This study used the same task types and multidimensional assessment of workload as Study 1. Due to the expertise and experience of the former NPP operators, the simplified simulator environment was restored to its original complexity of a full-scope NPP MCR simulator.

4.1 Overview

Former NPP operators employed by the NRC completed three basic control room operating tasks in a simulated environment. The three tasks were checking, detection, and response implementation presented on three touchscreen simulated panels. Associated performance and workload levels and types were evaluated. Performance measures included verifiable actions on the touchscreens, communication reporting accuracy, and navigation accuracy. Verifiable actions are any interactions with the interface. Workload was assessed using subjective measures including the NASA-TLX, MRQ, and ISA and using physiological responses recorded by EEG, fNIRS, TCD, and ECG. Operator participants were randomly assigned to either the RO1 or RO2 role. The role of the SRO was played by an experimenter. Results indicated a difference in workload and performance based upon role, which was associated with slightly different actions taken on controls on the panels. The detection task was the most challenging of the three tasks regardless of operator role.

4.2 Research Questions

The broader research goal of the HPTF research program was understanding operator performance in terms of what operators experience during each operationally relevant task. The primary and supplementary research questions from Study 1 included (For additional detail see section 3.2):

Primary Research Questions

- Can novice participants perform proficiently on realistic operator tasks?
- Were there differences in the level of proficiency achieved across the three task types (checking, detection, response implementation)?
- What are the workload levels and types associated with various types of tasks?
- What types of errors are associated with various task types?
- What workload measures are more sensitive and diagnostic to which types of tasks?

Supplementary Research Questions

- What are the types and levels of workload associated with each interface design?
- Is there an interaction between workload, display design, and task type?

With the exception of the first primary question which focused on discovering if novices could perform operator tasks proficiently, the research questions for Study 2 were similar to those of Study 1 (See section 3.2). Study 2 aimed to determine whether comparable workload findings would be obtained from an expert sample, relative to the results from novice samples in Study 1, which supports establishing the validity of the *different but equal approach*. Additionally, Study 2 enabled supplementary research questions related to operator role as both RO1 and RO2 roles were performed by experimental participants (in juxtaposition to the RO2 role being played by a confederate researcher in Study 1). Taken together, these methodological steps

and these research questions move us closer to an understanding of operator experience performing common main control room tasks.

4.3 Method

4.3.1 Participants

Participants for this experiment were formerly licensed NPP ROs, SROs, or navy/nuclear operators employed at the time of the experiment by the NRC. A demographics questionnaire was used to gather information about age, sex, PWR experience, and BWR experience. Eighteen (14 males, 4 females, *M = 45.94, SD =10.63*) participants were recruited through an agency announcement that was distributed to NRC employees. Participants had operational experience working in an MCR from the commercial nuclear power generation and/or naval nuclear power generation domains. These participants had experience with PWR and/or BWR technologies (Table 4-1 and 4-2). Participants were compensated for their participation by their regular hourly wage at the NRC.

Table 4-1. Number of participants claiming each type of experience.

Type of Experience	**PWR**	**BRW**	**Nuclear Navy**
Number of Participants	16	8	7

Table 4-2. Percentage of participants claiming one or more types of experiences.

Number of Experiences	**1**	**2**	**3**
Percentage	50%	28%	22%

4.3.2 Training Participants

Since the participants for this study were formerly licensed operators the training requirements were different in study 2 relative to study 1. Participants in Study 2 went through a two-hour training session conducted using PowerPoint and a GPWR NPP MCR simulator. The training session was used to ensure all participants were familiarized with the expected procedures and configuration of the GPWR NPP MCR simulator used in the experiment because each operational NPP MCR is different. The training covered three-way communication with the specific lexicon for the GPWR NPP MCR simulator, navigation within the simulated environment, and manipulation of I&Cs using the touchscreen interface. Participants in Study 2 did not have to meet the 80% performance requirement that was applied in Study 1 because of their baseline expertise in the nuclear domain.

4.3.3 Equipment

Experimental sessions were conducted in a laboratory room, set up as a mock MCR at the NRC headquarters in Rockville, Maryland. A GPWR NPP MCR simulator was configured for a crew of three operators. Crews consisted of two ROs and an SRO, whereby participants operated in the

role of either RO1 or RO2 and the researcher performed the role of SRO. The names RO1, RO2, and SRO were used to refer to the crew members in all communications.

4.3.3.1 Simulator

The GSE GPWR simulator was adapted for the present experiment. Simulator hardware consisted of four identical workstation computers connected locally using a gigabit network backbone. The hardware specifications for the workstation computers were Xeon X5650 6 core processor with a GeForce GTX 970 graphics card. In addition to the workstations, Microsoft Kinects and webcams were used to record communication events and video of the experimental conditions.

4.3.3.2 Interface

Each participant operated on two simulated control room wall panels. Panels consisted of four 27in monitors arranged in a two-by-two grid. Each 27in monitor had a resolution of 2560 pixels by 1440 pixels.

Figure 4-1 RO1 and RO2 operating on simulated control room wall panels

4.3.3.3 Physiological Instruments

A suite of three physiological instruments (Advanced Brain Monitoring's B-Alert X10, Spencer Technologies' ST3 Digital Transcranial Doppler, and Somantics' Invos Cerebral/Somatic Oximeter) was used to monitor workload states during experimental sessions for RO1. For RO2, workload state was monitored using Advanced Brain Monitoring's B-Alert X10. Details about the specific signals these systems monitored are explained in section 2.5.2.4.

4.3.3.4 Scenario Setup – Experimental Scenario

The same experimental scenario from Study 1 was used in Study 2, but with the complexity of the naming convention added back to nomenclature that would be meaningful for the formerly licensed population. As a review, the experimental scenario consisted of tasks from common steps required when completing operating procedures. The experimental scenario was developed based on a generic version for a "Loss of All Alternating Current Power" EOP as associated with the GSE GPWR simulator known generically as EOP-EPP-001. EOP-EPP-001

was the foundation for the simulator's initial condition but modified for experimental use. However, to maintain experimental control, other realistic tasks provided by an SME were incorporated. The GSE simulator physics information was used to determine when gauges, lights, temperature, pressure, etc. were at specific readings throughout EOP-EPP-001. Once that information was derived, the GSE simulator was adapted to strip away the physics-based functionality to allow for experimental control, ensuring each participant received the exact same experience for each condition. That experimental control is essential for drawing causal relation conclusions.

The modified EOP provided a narrative or context by which participants operated. Specifically, it required participants to perform predetermined tasks to respond to a loss of all alternating current power to the plant's safety buses (GSE Power Systems, 2011). The modified EOP required participant teams to utilize three control panels (A2, B1, C1) instead of four that would be required to execute the full EOP associated with EOP-EPP-001.

4.3.4 Experimental Scenario

In Study 1, to eliminate the need for operational knowledge and focus on the skill-based tasks, the number of I&Cs in each panel was reduced and the naming convention was simplified. Due to the experience and expertise of the former operators, the complexity that had been removed previously for the novice participants in Study 1 (described in Section 3.3.6) was re-introduced for study 2s (see Table 3-1 A2 Panel modification calculation and Figure 3-4 Original A2 panel used by operators (left) and modified A2 for experimentation) and naming convention and labeling used for the I&C (See Figure 3-5 Example of I&C name and alphanumeric code and Figure 3-6 Example of recoding I&C alphanumeric code of greater than seven digits) to ensure the ecological validity. By re-introducing the visual complexity representing real MCRs, the simulator could elicit comparable cognitive demand to former operator participants.

4.3.5 Experimental Design

Two randomly assigned participants were paired as a team for each experimental session. One participant performed the duties assigned to the role of RO1 and the other performed the duties assigned to the role of RO2. A repeated measure ANOVA for task type (checking, detection, and response implementation) and operator role (RO1, RO2) was employed in the present experiment.

NPP MCR procedures required performing a check of the I&C's state before executing a response implementation on the I&C. Therefore, task yoking was observed to maintain external validity. Each experimental session was randomly assigned one of three partially counterbalanced presentation orders for the pair of ROs as described in Table 4-1 Partial counterbalanced presentation order of tasks.

4.3.5.1 Dependent Measures

The same performance-based (task execution, communication), subjective, and physiological measures of workload used in Study 1 (see Section 3.3.8 Experimental Design) were used for Study 2. There was one small modification to the audio prompt used for the ISA subjective workload measure. The audio prompt contained the phrase, "RO1 [RO2] please rate your workload" as both roles were now participants.

Table 4-3 Partial counterbalanced presentation order of tasks

Presentation order	Task types		
Condition 1	Checking	Response implementation	Detection
Condition 2	Detection	Checking	Response implementation
Condition 3	Checking	Detection	Response implementation

In each experimental session a total of twenty-four steps were performed (eight checking steps, eight detection steps, and eight response implementation steps). To maintain equitable tasking, four of the eight steps within each condition were performed by RO1 and the remaining four were performed by RO2. Each step was independent with respect to functional fidelity such that actions performed during a step did not impact any of the other steps. The tasks alternated between the ROs such that RO1 always performs steps 1a, 2a, 3a, 4a and RO2 always performs steps 1b, 2b, 3b, 4b. To ensure ecological validity with the EOP, the eight steps within each task type were always performed sequentially in the identical order across sessions.

4.3.5.2 Independent Variable

The independent variables in this experiment were task type (checking, detection, and response implementation) and RO role (RO1, RO2).

4.3.5.2.1 Task Type

The task type consisted of the same three conditions as Study 1. The checking task type required a one-time inspection of an I&C to verify that it was in the state that the EOP called for it to be. Participants were required to locate various I&Cs and indicate identification by clicking on the correct I&C. The detection task type required participants to correctly locate an instrument then continuously monitor that instrument parameter for identification of change. Participants were required to monitor the instrument for five minutes and detect changes in level by clicking on a button located at the bottom of the instrument. Twelve random changes per minute occurred, totaling 60 changes per detection step. The response implementation task type required participants to locate a control and subsequently manipulate the control in the required direction (i.e., open or shut). Each task type consisted of four steps that were executed using three-way communication led by the experimenter acting as the SRO.

4.3.5.2.2 RO Role

The same tasks and task blocking were used in Study 1 as Study 2 and were distributed between the RO1 and RO2 roles similarly. For Study 2, however, former operator participants were placed in the RO2 role instead of a confederate participant as described in Section 3.3.4.

4.3.6 Procedure

The participants were provided an informed consent. Once finished reading the consent form, a demographics survey and pre-DSSQ questionnaire were administered. Since each NPP has slightly different operating procedures, training was conducted on the specific requirements for this simulated PWR NPP MCR. Training for all participants was conducted with the same researcher using PowerPoint and the EPIC simulator. The training lasted around two hours. At the conclusion of the training, participants were given a five-minute break.

After the break, the RO1 and RO2 participants were connected to physiological sensors. RO1 was connected to fNIRS, TCD, EEG, and ECG sensors. RO2 was connected to EEG and ECG sensors. Once both participants were connected to the sensors, five-minute wakeful rest baseline data were collected simultaneously on the two paired participants. After the baseline, the three experimental conditions were performed sequentially (in the presentation order randomly assigned for the participant team). The SRO used three-way communication during each condition to initiate the eight tasks. The SRO alternated tasking between RO1 and RO2 such that RO1 performs tasks 1a, 2a, 3a, 4a and RO2 performs tasks 1b, 2b, 3b, 4b. An audible ISA prompt was automatically triggered halfway through tasking for each RO (e.g., immediately following the completion of task 2a, the computer played an audio prompt saying "RO1 please rate your workload"). Upon completion of each task type, both participants were given the NASA-TLX, MRQ, and DSSQ questionnaires electronically. Lastly, upon completion of the third task type and questionnaires, all physiological sensors were removed, and the participants were debriefed and dismissed.

4.4 Results

4.4.1 Workload

ANOVAs of subjective and objective metrics were used to determine if there was a significant difference between workload experienced during the three different task types (checking, detection, and response implementation), the two different RO roles (RO1 and RO2), and the interaction between these two factors. Greenhouse-Geisser corrections were used where the assumption of sphericity was not met and, to account for Type I errors, Bonferroni corrections were used for post-hoc comparisons. The Brown-Forsythe[17] statistic is reported when homogeneity of variance is violated for one-way ANOVAs.

4.4.1.1 Subjective Workload Measures

4.4.1.1.1 NASA-TLX

A 3 (task type*:* checking, detection, and response implementation) × 2 (RO role: RO1 and RO2) mixed ANOVA with repeated measures on task types was conducted for each of the subscales. The ANOVA was used to determine if there was a significant perceived workload difference between task types, types of workloads, and roles. The analysis would also reveal if task type effects differed for the two RO roles, and if different combinations of task and RO roles elicited different patterns of workload response, as tapped by the NASA-TLX subscales.

A significant main effect was found for task type, F (2, 32) = 8.61, p < .01, η_p^2 = .35 such that participants perceived greater workload during detection (M = 37.55, SD = 22.57) compared to the checking (M = 23.70, SD = 17.70) and response implementation (M = 25.00, SD = 18.06) task types (Figure 4-2 Global NASA-TLX means by task type (error bars denote standard errors)).

17 The Brown–Forsythe test statistic is the F statistic resulting from an ordinary one-way analysis of variance on the absolute deviations of the groups or treatments data from their individual medians.

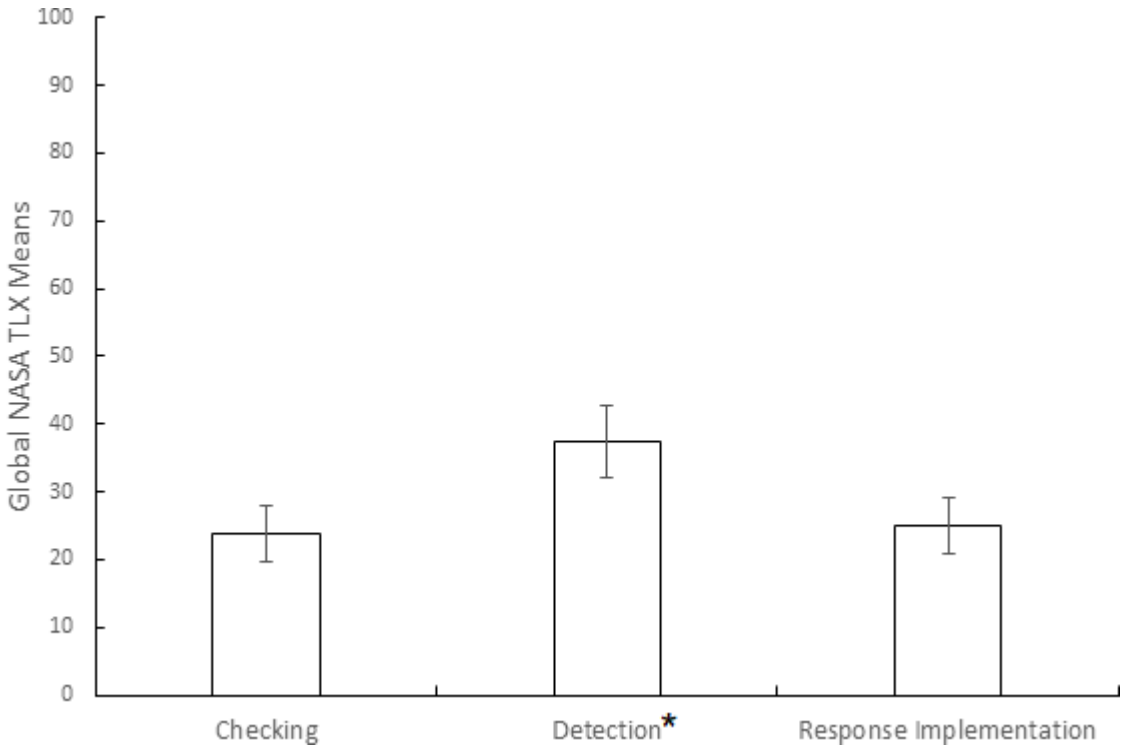


Figure 4-2 Global NASA-TLX means by task type (error bars denote standard errors)

A significant main effect was found for the subscales of the NASA-TLX, F (2.76, 44.15) = 6.40, p < .01,η_p^2 = .29, such that participants reported mental demand (M = 38.33, SD = 23.30) higher than physical demand (M = 18.61, SD = 15.71) temporal demand (M = 22.41, SD = 16.15) and effort (M = 30.83, SD = 21.84) ratings (Figure 4-3 NASA-TLX scores by subscale (error bars denote standard errors)).

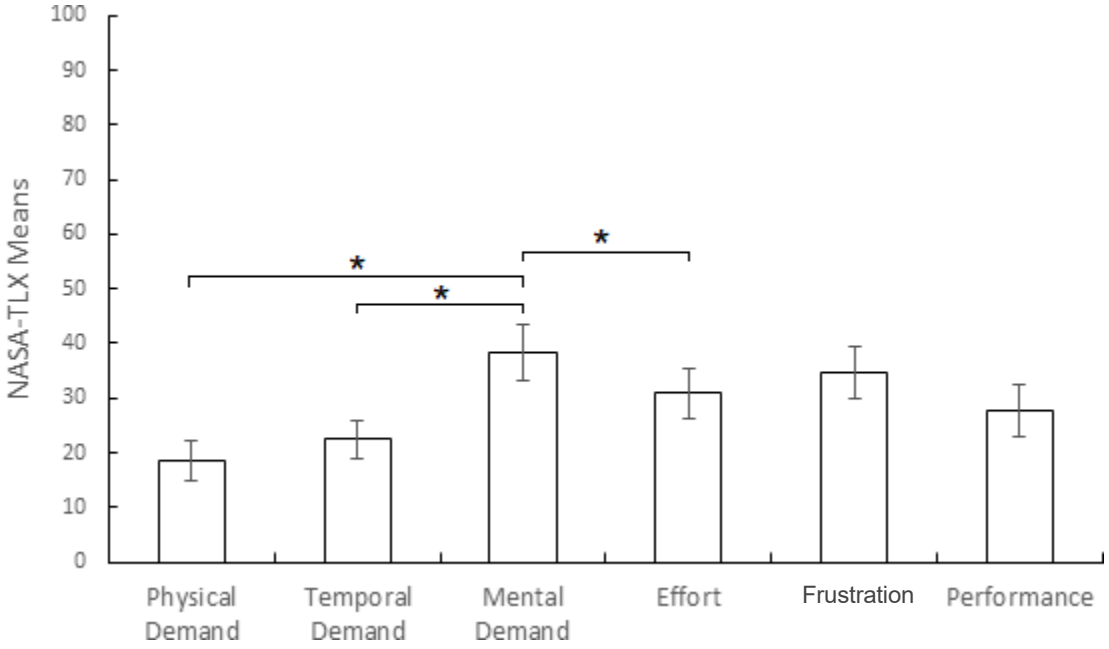


Figure 4-3 NASA-TLX scores by subscale (error bars denote standard errors)

A significant main effect for role was found, F (1, 16) = 6.79, p = .02, η_p^2 = .30, such that workload ratings were generally higher for RO2 (M = 37.93, SD = 16.17) compared to RO1 (M = 19.57, SD = 13.61) No significant interactions were found (p > .05) (Figure 4-4 Global NASA-TLX means by RO role (error bars denote standard errors)).

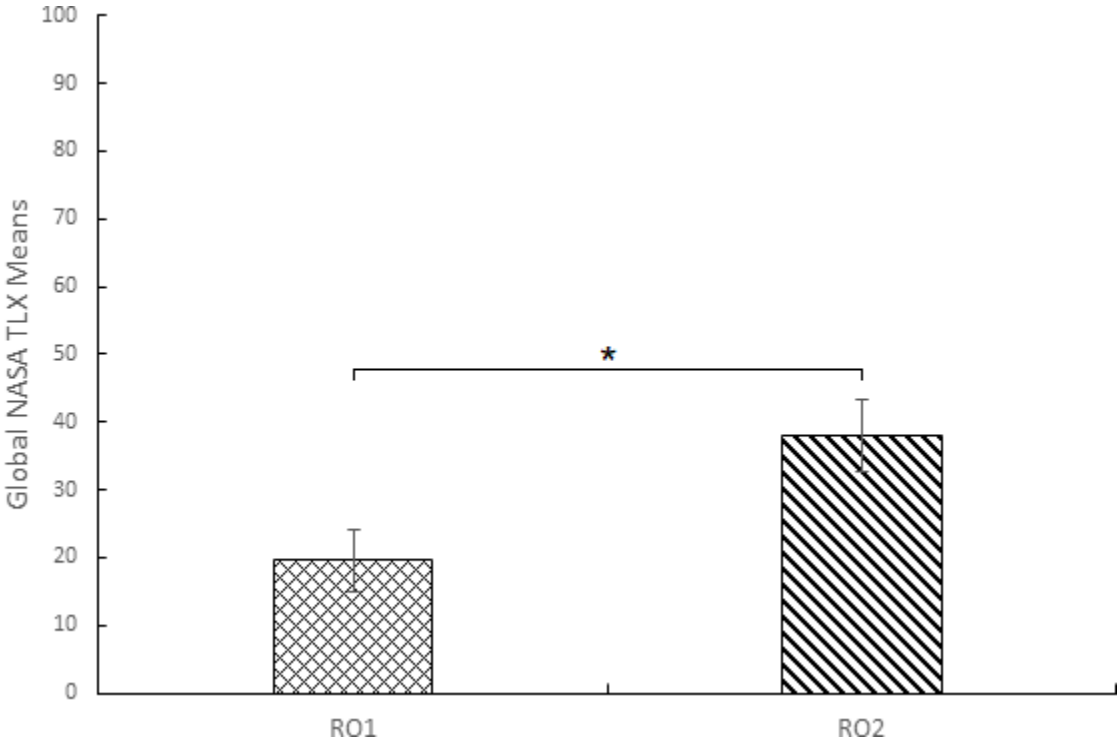


Figure 4-4 Global NASA-TLX means by RO role (error bars denote standard errors)

4.4.1.1.2 ISA

A 3 (task type*:* checking, detection, and response implementation*) × 2* (RO role*:* RO1 and RO2) mixed ANOVA with repeated measures on task type was conducted. The ANOVA was used to determine if task type and RO role have significant effects on online-subjective workload (i.e., reported as the tasks were being performed), and to determine if the pattern of workload differences found across the tasks differed between RO roles. For ISA ratings, no significant main effect was found for task type ($p > .05$) or RO role ($p > .05$). No interaction was found between task type and RO role for ISA ratings ($p > .05$)

4.4.1.1.3 MRQ

A 3 (task type: checking, detection, and response implementation) × 2 (RO role: RO1 and RO2) mixed ANOVA with repeated measures on task type was conducted for each of the fourteen MRQ subscales tapping various processes that contribute to the workload experienced. The ANOVAs would reveal if the task types and RO roles had any overall effects on workload from the activation of the processes as assessed by the MRQ. The analyses would also show whether the effects of task type on the workload ratings were different or consistent between the two RO roles.

Task type showed a significant effect on several of the subscales that persisted across the RO roles. A significant main effect was found for task type for manual subscale, $F(2, 32) = 11.18$, $p < .01$, $\eta_p^2 = .41$, such that detection ($M = 66.83$, $SD = 17.22$) yielded a higher rating than checking ($M = 46.11$, $SD = 23.29$) and response implementation ($M = 55.11$, $SD = 26.16$) did not differ from either checking or detection. A significant main effect was found for task type for spatial attentive subscale, $F(1.49, 23.78) = 5.21$, $p = .02$, $\eta_p^2 = .25$, such that detection ($M =$

75.72, *SD* = 15.00) yielded a higher rating than checking (*M* = 63.06, *SD* = 19.66) and response implementation (*M* = 63.94, *SD* = 21.73) did not differ from either checking or detection. A significant main effect was found for task type for spatial concentrative subscale, *F*(1.43, 22.88) = 7.05, $p < .01$, $\eta_p^2 = .31$, such that detection (*M* = 66.61, *SD* = 25.99) was higher than both checking (*M* = 41.00, *SD* = 26.82) and response implementation (*M* = 43.83, *SD* = 30.56) A significant main effect was found for task type for spatial quantitative subscale, F(2, 32) = 25.44, $p < .01$, $\eta_p^2 = .61$, such that detection (*M* = 66.94, *SD* = 27.13) was higher than both checking (*M* = 23.67, *SD* = 22.54) and response implementation (*M* = 25.89, *SD* = 28.83).

A significant main effect was found for task type for the visual temporal subscale, *F*(2, 32) = 4.98, $p = .03$, $\eta_p^2 = .24$, but pairwise comparisons did not show any differences between checking (*M* = 18.50, *SD* = 22.87) detection (*M* = 36.17, *SD* = 30.52), and response implementation (*M* = 22.00, *SD* = 24.69). A significant main effect was found for task type for the vocal process subscale, *F*(2, 32) = 4.46, $p = .02$, $\eta_p^2 = .22$, but pairwise comparisons did not show differences between checking (*M* = 63.83, *SD* = 18.50) detection (*M* =50.78, *SD* = 20.23), and response implementation (*M* = 57.50, *SD* = 25.40).

A significant main effect was found for RO role for the spatial attentive subscale, *F*(1, 16) = 5.59, $p = .03$, $\eta_p^2 = .26$, such that RO1 (*M* = 59.81, *SD* = 11.18) participants rated lower on the spatial attentive subscale compared to RO2 (*M* = 75.33, *SD* = 16.21) participants.

The effects on the RO role for spatial emergent processing differed for different tasks. A significant interaction effect was found between task type and RO role for the spatial emergent sub-scale, *F*(2, 32) = 3.96, $p = .03$, $\eta_p^2 = .20$, such that the response implementation task elicited lower ratings from RO1 participants (*M* = 53.11, *SD* = 13.20) than RO2 participants (*M* = 75.44, *SD* = 15.54), but ratings were not significantly different between RO roles for the checking and detection task types. (Figure 4-5 MRQ Spatial Emergent scores by task type and RO role (error bars denote standard errors)).

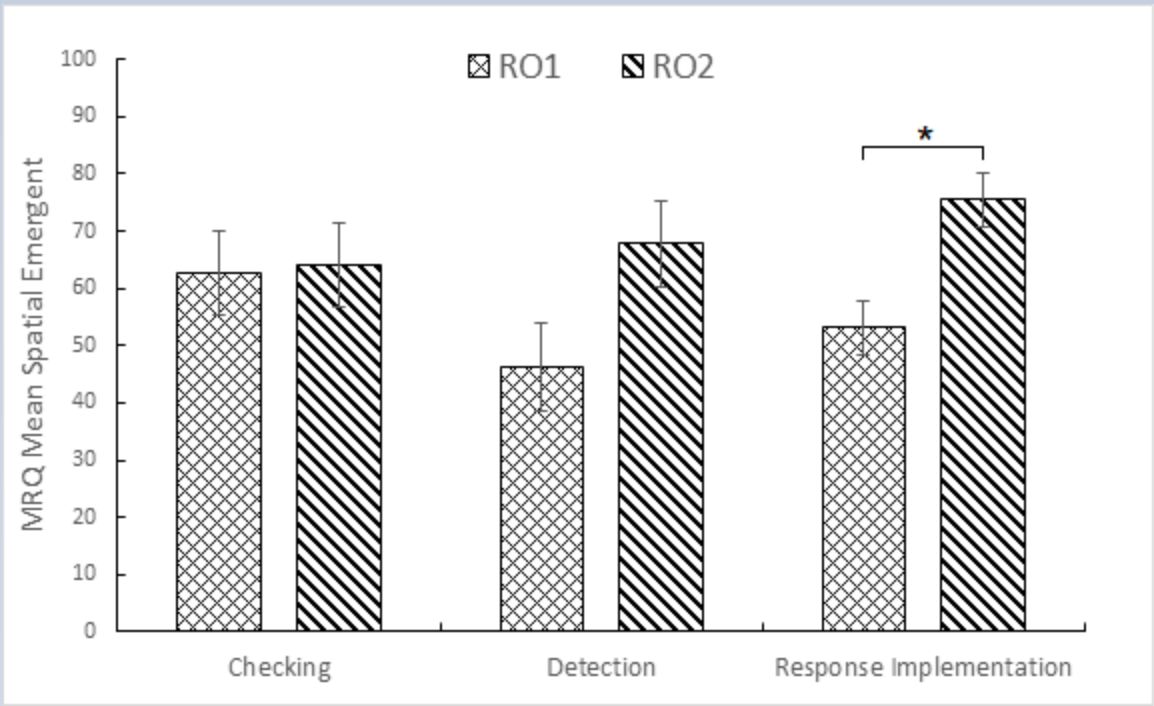


Figure 4-5 MRQ Spatial Emergent scores by task type and RO role (error bars denote standard errors)

4.4.2 Physiological Measures

As in study 1, all physiological dependent variables were calculated as differences from a five-minute resting baseline. For example, if the participant's left CBFV for the five-minute baseline was 73.23 cm/s and their left CBFV for the subsequent checking task was 75.33 cm/s, their difference from baseline would be 2.10 cm/s. This method helps account for individual differences when comparing group means as is the case when running ANOVAs.

4.4.2.1 Electroencephalogram (EEG)

Brain activity was recorded at 9 EEG sensor sites: the EEG data were analyzed by grouping sensor sites by hemispheres (i.e., compare brain activity between the left and right hemispheres) and lobes (i.e., compare brain activity among the frontal, parietal and occipital lobes).

4.4.2.1.1 Hemispheres

A 3 (task type: checking, detection, and response implementation) × 2 (RO role: RO1 and RO2) mixed ANOVA was run for left and right hemispheres separately. These ANOVAs provided insight into the overall effects of task type and RO role on activity in the left and right hemispheres. ANOVAs were run separately or theta, alpha, and beta frequency bands.

Theta Waves for the Left and Right Hemispheres

For left hemisphere theta, a significant main effect was found for task type, $F(2, 30) = 6.03$, $p < .01$, $\eta_p^2 = .29$, such that detection ($M = -109.91$, $SD = 1404.91$) elicited lower theta than

checking (*M* = 767.49, *SD* = 1836.26) and response implementation (*M* = 778.09, *SD* = 2076.29) did not differ from either checking or detection (Figure 4-6 Theta left hemisphere change from baseline in μV2 by task type (error bars denote standard errors)). No significant main effect for RO role and no interaction was found for left hemisphere theta ($p > .05$).

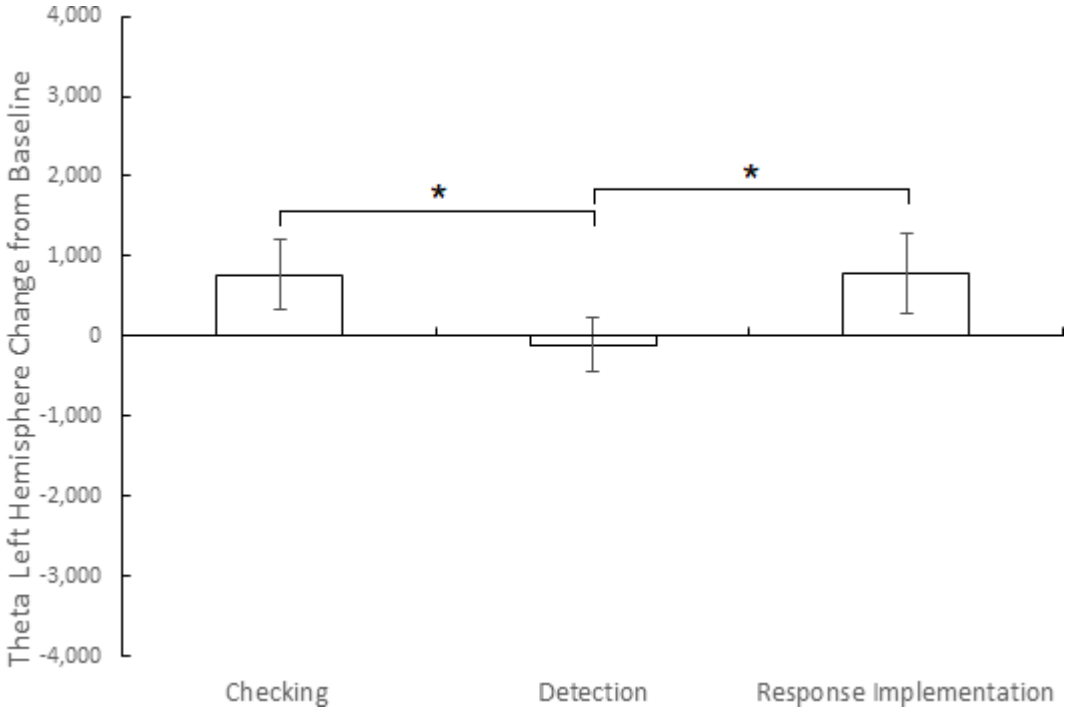


Figure 4-6 Theta left hemisphere change from baseline in μV2 by task type (error bars denote standard errors)

For right hemisphere theta, a significant main effect was found for task type, $F(2, 30) = 7.38$, $p < .01$, $\eta_p^2 = .33$, such that detection (*M* = 151.51, *SD* = 877.90) elicited lower theta than both checking (*M* = 594.47, *SD* = 1236.31) and response implementation (*M* = 570.64, *SD* = 1164.32) (Figure 4-7 Theta right hemisphere change from baseline in μV2 by task type (error bars denote standard errors)). No significant main effect for RO role and no interaction was found for left hemisphere theta ($p > .05$).

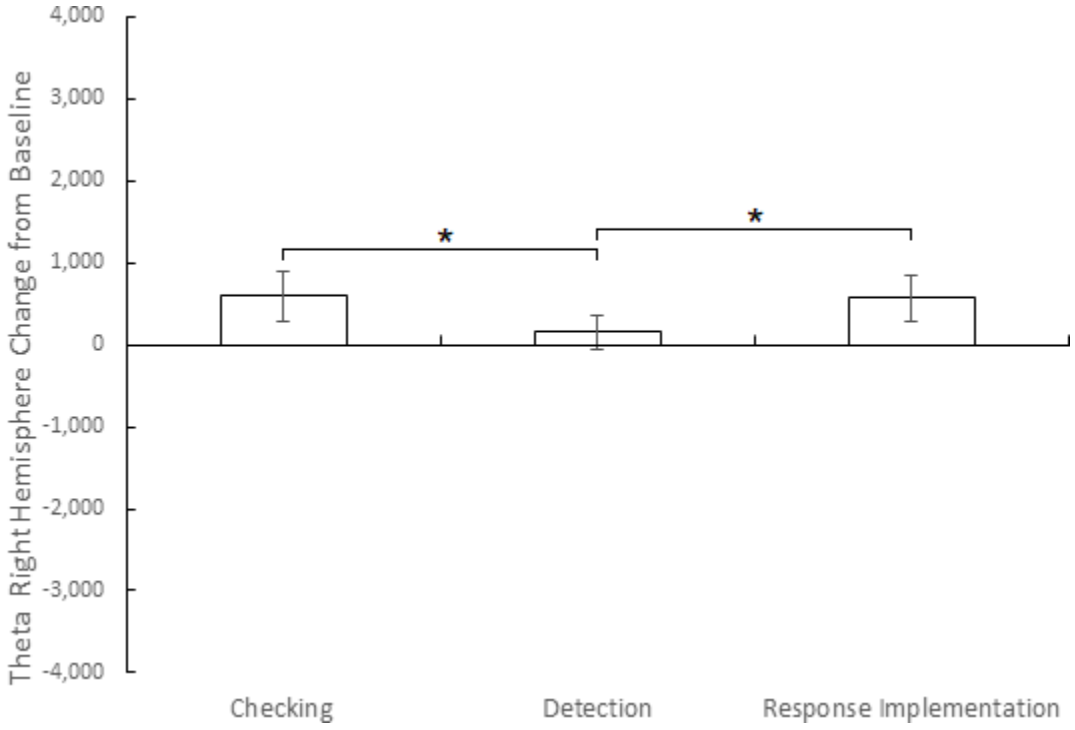


Figure 4-7 Theta right hemisphere change from baseline in μV2 by task type (error bars denote standard errors)

Alpha Waves for the Left and Right Hemispheres

For left hemisphere alpha, a significant main effect was found for task type, $F(2, 30) = 3.56$, $p = .04$, $\eta_p^2 = .19$, such that detection ($M = -994.77$, $SD = 1941.13$) alpha had a greater decrease from baseline in the left hemisphere than response implementation ($M = -536.11$, $SD = 1981.41$), but checking ($M = -434.51$, $SD = 2438.28$) did not differ from either detection or response implementation. No main effect for RO role was found for left hemisphere alpha ($p > .05$). The interaction effect between task type and RO role for alpha in the left hemisphere was at the threshold for statistical significance, $F(2, 30) = 3.45$, $p = .05$, $\eta_p^2 = .19$, so not surprisingly, pairwise comparison did not yield any significant results (Figure 4-8 Alpha right hemisphere change from baseline in μV2 by task type (error bars denote standard errors)).

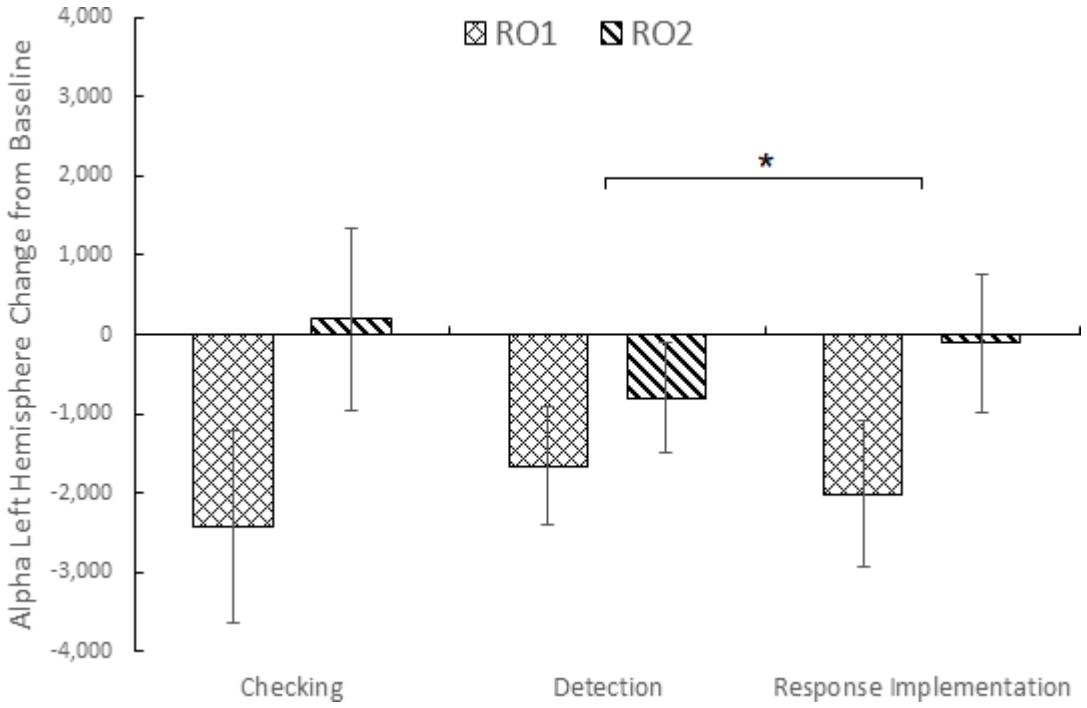


Figure 4-8 Alpha right hemisphere change from baseline in µV2 by task type (error bars denote standard errors)

For the right hemisphere alpha, no significant effects were found ($p > .05$).

Beta Waves for the Left and Right Hemispheres

For left hemisphere beta, no significant effects were found ($p > .05$). For right hemisphere beta, a significant main effect was found for task type, $F(2, 30) = 7.95$, $p < .01$, $\eta_p^2 = .35$, such that detection ($M = 314.69$, $SD = 1899.50$) elicited lower beta than both checking ($M = 1518.30$, $SD = 2192.64$) and response implementation ($M = 1670.59$, $SD = 1983.88$). No significant main effect for RO role was found for beta in the right hemisphere ($p > .05$). A significant interaction effect was found between task type and RO role for right hemisphere beta, $F(2, 30) = 5.28$, $p = .01$, $\eta_p^2 = .26$. However, the simple effects did not reveal any significant differences in beta between RO1 and RO2 for any of the task types (Figure 4-9 Beta right hemisphere change from baseline in µV2 by task type and RO role (error bars denote standard errors)).

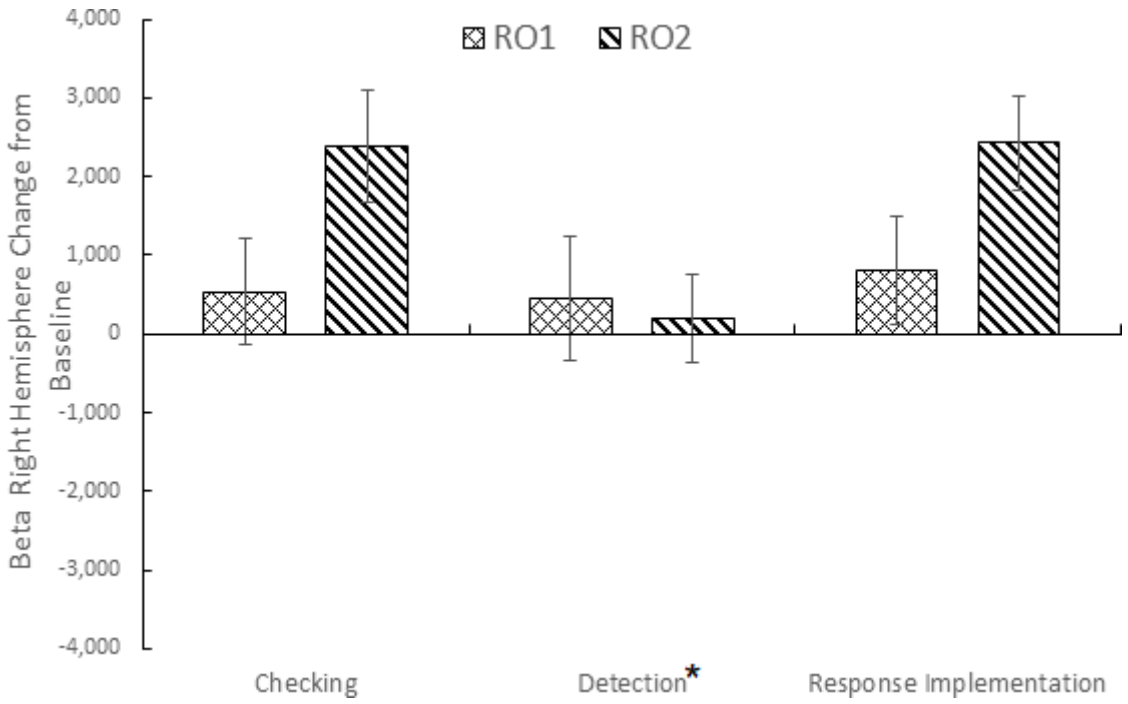


Figure 4-9 Beta right hemisphere change from baseline in μV2 by task type and RO role (error bars denote standard errors)

4.4.2.1.2 Lobes

A 3 (task type: checking, detection, and response implementation) × 2 (RO role: RO1 and RO2) mixed ANOVA was run for frontal, parietal, and occipital lobes separately. These ANOVAs provided insight into the overall effects of task type and RO role on the frontal, parietal, and occipital lobes. ANOVAs were run separately for theta, alpha, and beta frequency bands.

Theta for the Frontal, Parietal, and Occipital Lobes

For frontal lobe theta, no effects were found ($p > .05$). For the parietal lobe theta, a significant main effect was found for task type, $F(2, 30) = 7.42$, $p < .01$, $\eta_p^2 = .33$, such that detection ($M = -17.99$, $SD = 1033.98$) was lower than both checking ($M = 691.32$, $SD = 1413.26$) and response implementation ($M = 499.03$, $SD = 1408.24$) (Figure 4-10 Theta parietal lobe change from baseline in μV2 by task type (error bars denote standard errors)). For the occipital lobe theta, no effects were found ($p > .05$).

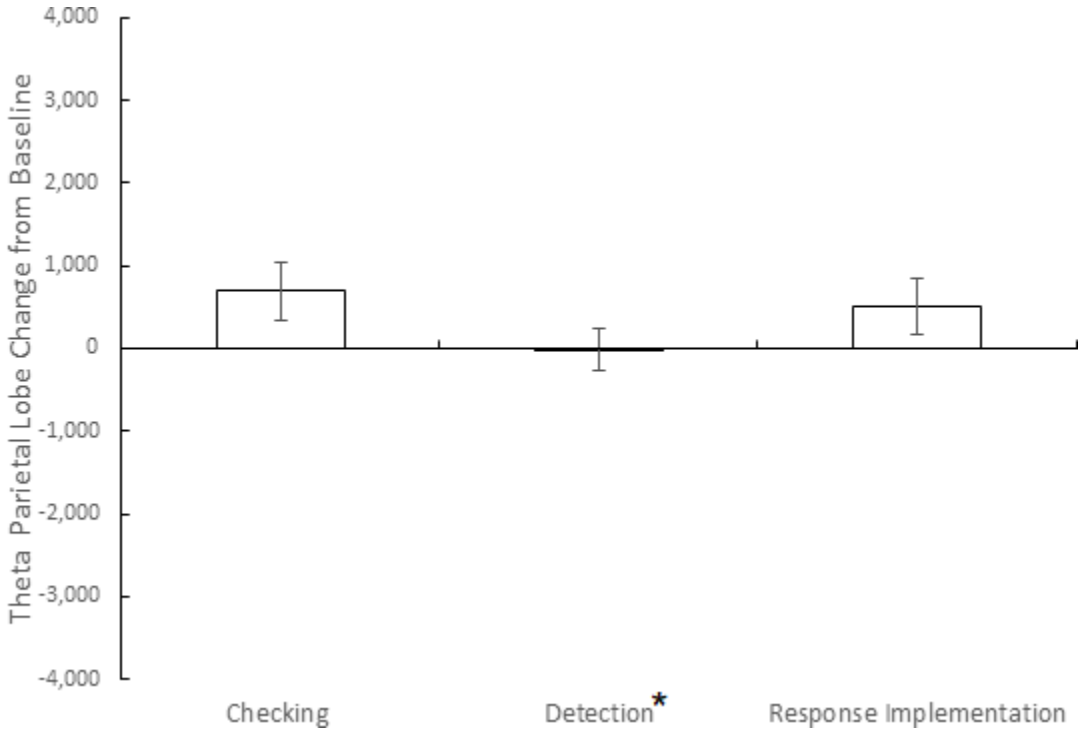


Figure 4-10 Theta parietal lobe change from baseline in μV2 by task type (error bars denote standard errors)

Alpha for the Frontal, Parietal, and Occipital Lobes

No significant effects were found for frontal, parietal, or occipital lobe alpha ($p > .05$).

Beta for the Frontal, Parietal, and Occipital Lobes

For frontal lobe beta, no statistically significant effects were found ($p > .05$). For the parietal lobe beta, a significant main effect was found for task type, $F(2, 30) = 3.68$, $p = .04$, $\eta_p^2 = .20$, but pairwise comparisons did not show a difference between checking ($M = 1926.20$, $SD = 1451.47$), detection ($M = 484.38$, $SD = 1837.55$) and response implementation ($M = 1457.99$, $SD = 1372.53$). No significant main effect for RO role and no interaction was found for parietal lobe beta ($p > .05$). For occipital lobe beta, a significant main effect was found for task type, $F(1.16, 17.44) = 6.25$, $p = .02$, $\eta_p^2 = .29$, such that detection ($M = 52.17$, $SD = 1581.02$) was lower than both checking ($M = 714.59$, $SD = 1536.53$) and response implementation ($M = 1245.79$, $SD = 2213.63$) (Figure 4-11 Beta occipital lobe change from baseline in μV2 by task type (error bars denote standard errors)). No main effect for RO role and no interaction was found for occipital lobe beta ($p > .05$).

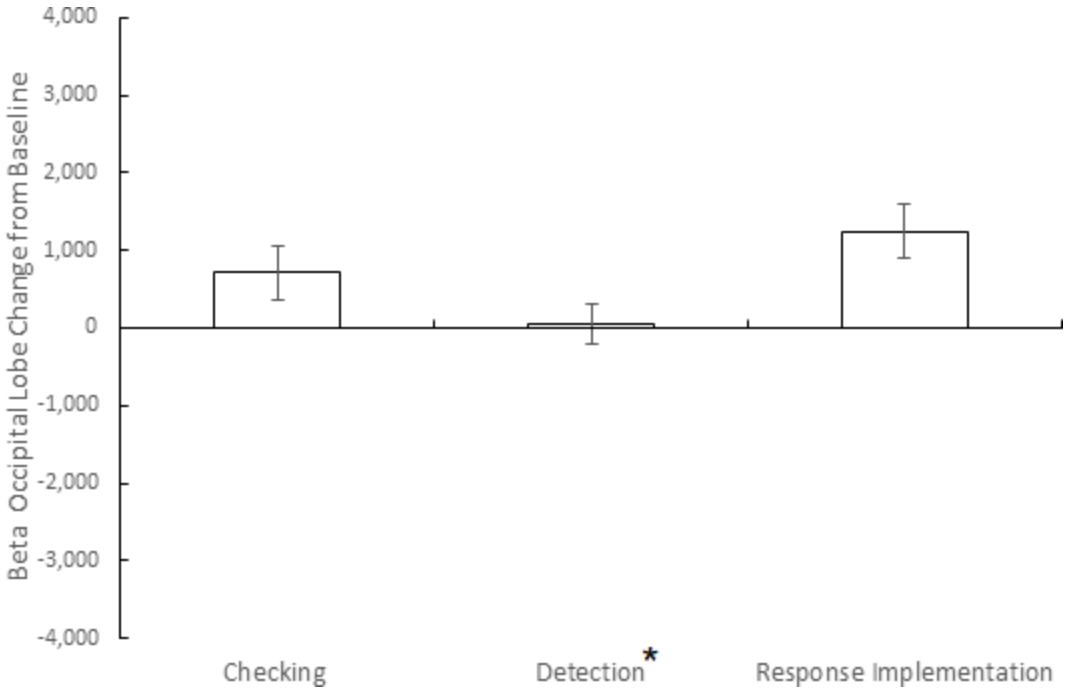


Figure 4-11 Beta occipital lobe change from baseline in μV2 by task type (error bars denote standard errors)

4.4.2.2 Transcranial Doppler Ultrasonography (TCD)

The TCD sensor was only available for data collection on RO1. Therefore, ANOVAs for this measure only included RO1 data. One-way ANOVAs (task type: checking, detection, and response implementation) were conducted to determine if there were overall effects of task type on mean CBFV. ANOVAs were run separately for CBFV in the left and right medial cerebral arteries. No significant effect was found for CBFV recorded on the left medial cerebral artery for RO1 participants by task type ($p > .05$). Also, no significant effect was found for CBFV right medial cerebral artery for RO1 participants by task type ($p > .05$).

4.4.2.3 Functional Near-Infrared Spectroscopy (fNIRS)

The fNIRS sensor was only available for data collection on RO1. Therefore, ANOVAs for this measure only included RO1 data. A one-way (task type: checking, detection, and response implementation) ANOVA was run to determine if regional oxygen saturation (rSO_2) was significantly different across task types. A fNIRS rSO_2 ANOVA was run separately for the left and right prefrontal cortex.

A significant effect was found for task type for left pre-frontal cortex rSO_2 for RO1 participants, $F(2, 16) = 5.41$, $p = .02$, $\eta_p^2 = .40$, such that detection ($M = 2.46$, $SD = 1.40$) was higher than response implementation ($M = 1.32$, $SD = 1.52$), and checking ($M = 1.45$, $SD = 1.55$) did not differ from either detection or response implementation (Figure 4-12 Left pre-frontal cortex rSO2 change from baseline for RO1 participants by task type (error bars denote standard errors)).

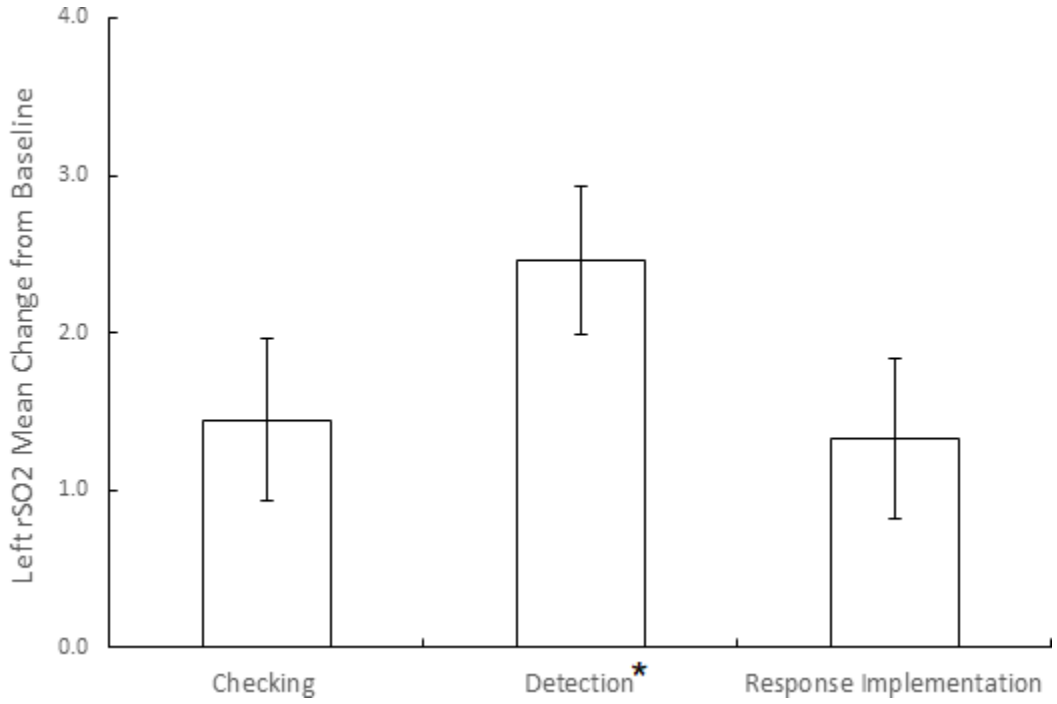


Figure 4-12 Left pre-frontal cortex rSO2 change from baseline for RO1 participants by task type (error bars denote standard errors)

A significant effect was found for task type for right pre-frontal cortex rSO_2 for RO1 participants, $F(2, 16) = 8.04$, $p < .01$, $\eta_p^2 = .57$, such that detection ($M = 2.46$, $SD = 1.40$) was higher than response implementation ($M = 1.32$, $SD = .52$), and checking ($M = 1.45$, $SD = 1.55$) did not differ from either detection or response implementation (Figure 4-13 Right pre-frontal cortex rSO2 change from baseline for RO1 participants by task type (error bars denote standard errors)).

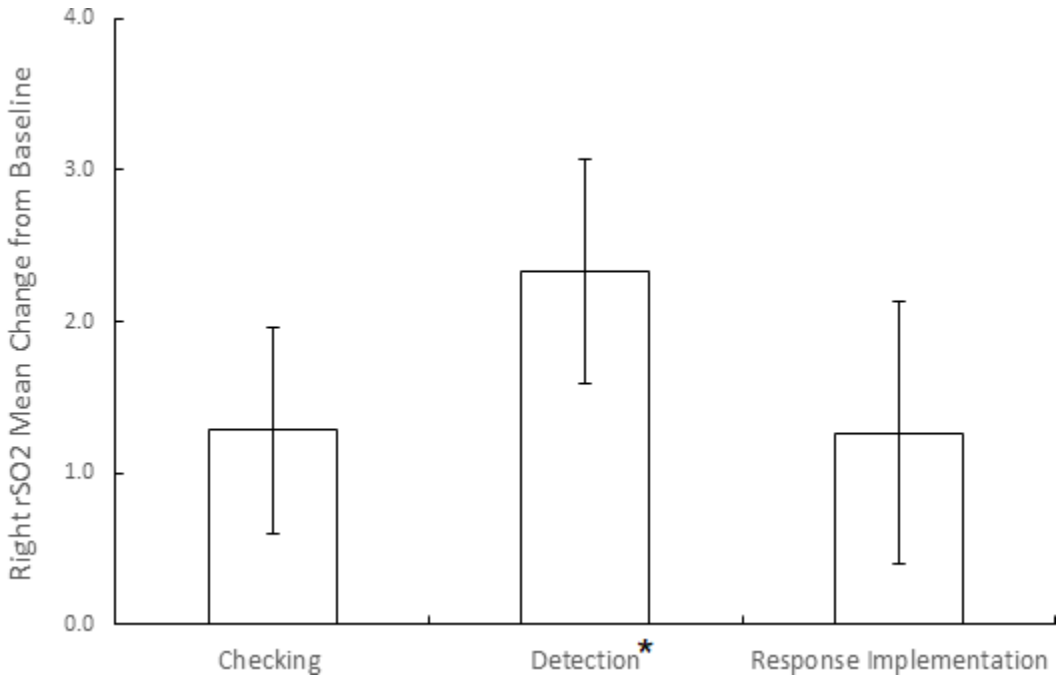


Figure 4-13 Right pre-frontal cortex rSO2 change from baseline for RO1 participants by task type (error bars denote standard errors)

4.4.2.4 Electrocardiogram (ECG)

Three (task type: checking, detection, and response implementation) × two (RO role: RO1 and RO2) mixed ANOVAs with repeated measures on task type were conducted to determine if the different task types and RO role affected HR, HRV, and IBI. These analyses also assessed the interaction between the task types and RO roles, which would reveal if any differences that occurred across task types were similar for the RO1 and RO2 participants. HR, IBI, and HRV were derived from R-Peak detections using the So-Chan QRS algorithm from the raw ECG signal. No significant effects were found for HR, IBI, or HRV ($p > .05$).

4.4.3 Performance Measures

4.4.3.1 Communication Reporting

Communication reporting variables included percent communications completed correctly, number of I&C location help requests, number of clarifications required, and number of requests for repeating an instruction. Four 3 (task type: checking, detection, and response implementation) × 2 (RO role: RO1 and RO2) mixed ANOVAs with repeated measures on task type were conducted for each of the four measures to determine if there was a significant difference between task types and between RO roles.

A significant main effect was found for task type for percentage of communications completed correctly, $F(2, 32) = 6.94$, $p < .01$, $\eta_p^2 = .30$, such that detection ($M = 66.67$, $SD = 32.93$) was significantly lower than response implementation ($M = 65.28$, $SD = 39.42$), and checking ($M = 65.28$, $SD = 39.42$) did not differ from either detection or response implementation. No significant main effect for RO role was found for percentage of communications completed

correctly (p > .05). A significant interaction effect was found between task type and RO role for percentage of communications completed correctly, $F(2, 32) = 4.25$, $p = .02$, $\eta_p^2 = .21$. However, the subsequent simple effects analyses did not yield significance between the percentage of communications completed correctly between RO1 and RO2 for each of the task types (Figure 4-14 Percentage of communications completed correctly by task type and RO role (error bars denote standard errors)).

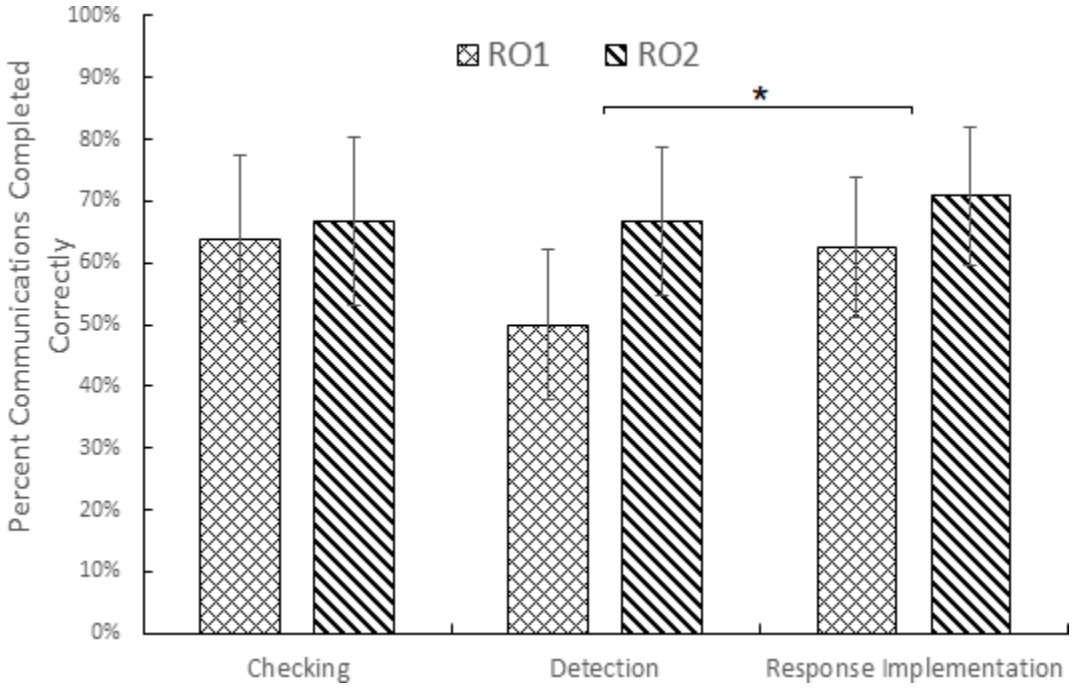


Figure 4-14 Percentage of communications completed correctly by task type and RO role (error bars denote standard errors)

No significant main effects or interaction ware found for the number of requests for location help (p > .05).

No significant main effects or interaction effects were found for the number of clarifications (p > .05).

For requests for repeating instructions, a significant main effect was found for task type, $F(2, 32) = 9.09$, $p < .01$, $\eta_p^2 = .36$, such that detection ($M = 1.56$, $SD = 1.25$) had more requests for repeating instructions than both checking ($M = 0.56$, $SD = 0.78$) and response implementation ($M = 0.67$, $SD = .14$) (Figure 4-15 Mean number of repeat instruction requests by task type (error bars denote standard errors)). No significant main effect or interaction was found for the number of requests for repeating instructions (p > .05).

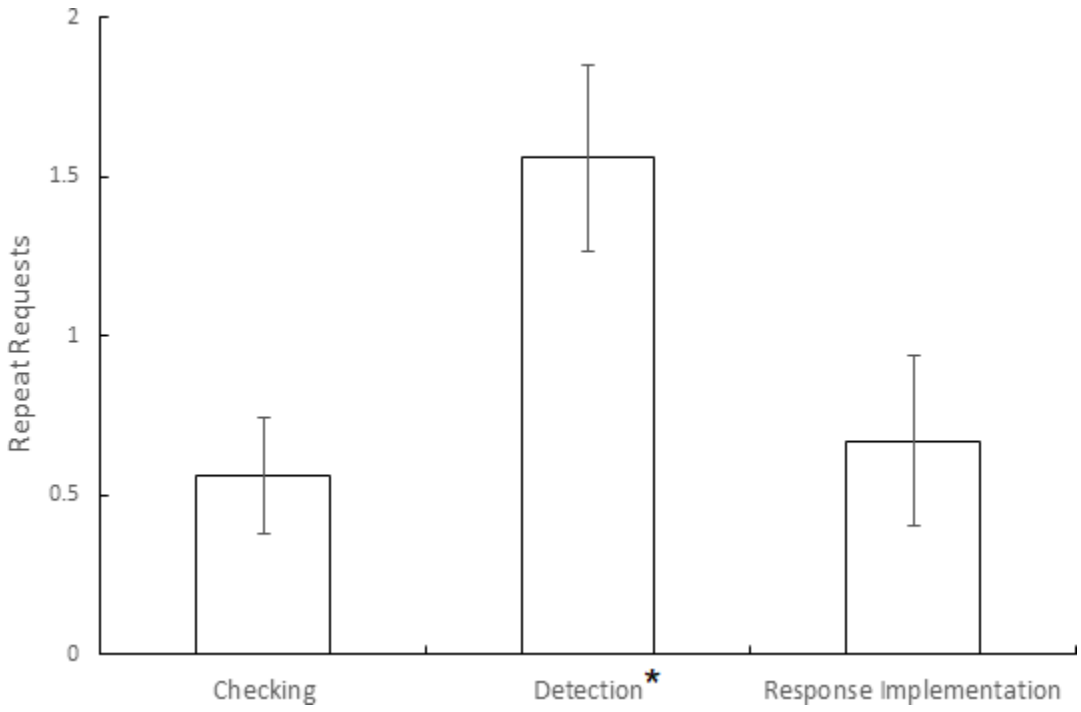


Figure 4-15 Mean number of repeat instruction requests by task type (error bars denote standard errors)

4.4.3.2 Navigation and Identification

Navigation and identification variables included (i) number of correctly identified I&Cs, (ii) locating and identifying the correct I&C on the first attempt, (iii) the number of additional identifications made on the correct I&C, and (iv) number of incorrect identifications. Four 3 (task type: checking, detection, and response implementation) × 2 (RO role: RO1 and RO2) mixed ANOVAs were conducted for each of the four measures to determine if there was a significant difference between task types and between RO role. The analyses also revealed if the RO1 and RO2 roles showed similar patterns of differences in performance across the task types. Task type was a repeated-measures variable and RO role was a between-subjects variable.

For number of correctly identified I&Cs, no significant main effect was found for task type ($p > .05$). A significant main effect was found for RO role for the number of correct identification actions, $F(1,16) = 5.26$, $p = .04$, $\eta_p^2 = .25$, such that RO1 ($M = 3.82$, $SD = .24$) participants correctly identified fewer I&Cs compared to RO2 ($M = 4$, $SD = 0$) participants. No interaction was found between the task type and RO role for the number of correct identification actions ($p > .05$).

For locating and identifying the correct I&C on the first attempt, no significant main effects or interactions were found ($p > .05$).

For the number of additional identifications made on the correct I&C, a significant main effect was found for task type, $F(1.01, 16.16) = 16.55$, $p < .01$, $\eta_p^2 = .51$, such that detection ($M = 12.89$, $SD = 12.87$) was higher than both checking ($M = .39$, $SD = .98$) and response

implementation (*M* = 0.28, *SD* = 0.46). There was no significant main effect for RO role or role x task type interaction for the number of additional identification actions ($p > .05$).

For the number of incorrect identifications, no significant main effect was found for task type ($p > .05$). There was a significant main effect for RO role, $F(1, 16) = 4.97$, $p = .04$, $\eta_p^2 = .24$, indicating that the number of incorrect identifications differed by RO role. RO1 (*M* = 0.04, *SD* = 0.11) participants had fewer incorrect identifications compared to RO2 (*M* = 0.26, *SD* = 0.28) participants (Figure 4-16 Mean number of additional identifications by task type (error bars denote standard errors)). No interaction effect was found between the task type and RO role for the number of incorrect identifications ($p > .05$).

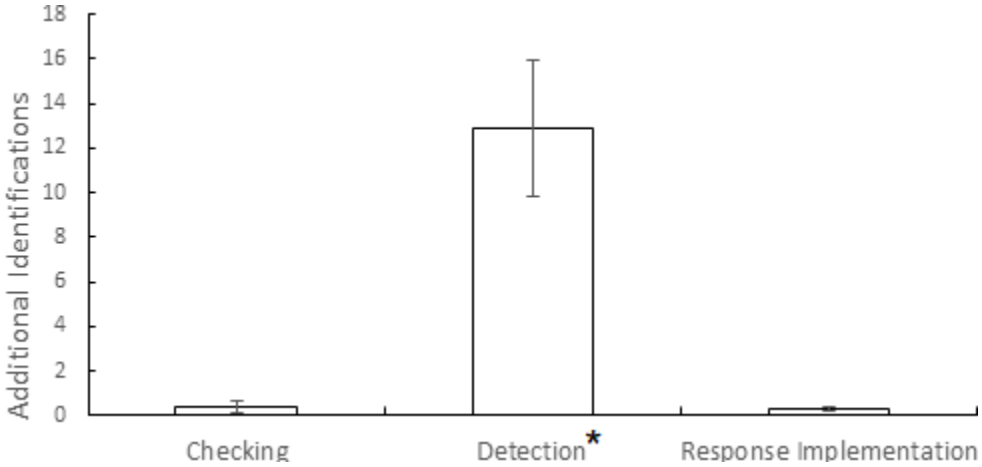


Figure 4-16 Mean number of additional identifications by task type (error bars denote standard errors)

4.4.3.3 Action

Independent sample t-tests were conducted to determine if there were significant differences between RO role for various action performance variables. Below are the descriptions and results of each action performance measure for detection and response implementation.

4.4.3.3.1 Detection

The percent of correct gauge change detections, percent of missed gauge change events, and the number of false positive detections for each participant were measured while completing the detection task. No difference in percent of correct gauge change detections was found for RO1 (*M* = 47.86, *SD* = 22.74) and RO2 (*M* = 33.21, *SD* = 23.51) roles; ($p > .05$) (Figure 4-17 Percent correct detections by RO role (error bars denote standard errors)). No difference in percent of missed gauge change events was found for RO1 (*M* = 43.50, *SD* = 23.23) and RO2 (*M* = 58.61, *SD* = 26.55) roles ($p > .05$) (Figure 4-18 Percent missed change events by RO role (error bars denote standard errors). No difference in the number of false positive detections was found for RO1 (*M* = 61.11, *SD* = 45.71) and RO2 (*M* = 78.78, *SD* = 42.32) roles ($p > .05$) (Figure 4-19 Number of false positive detections by RO role (error bars denote standard errors)).

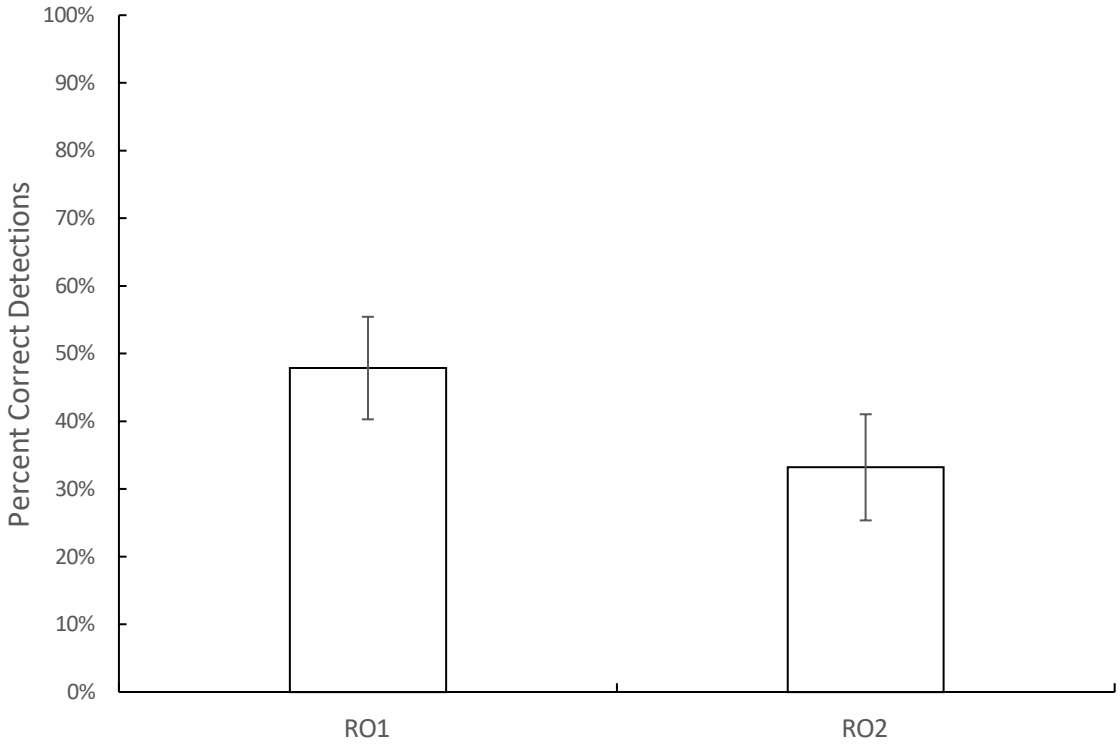


Figure 4-17 Percent correct detections by RO role (error bars denote standard errors)

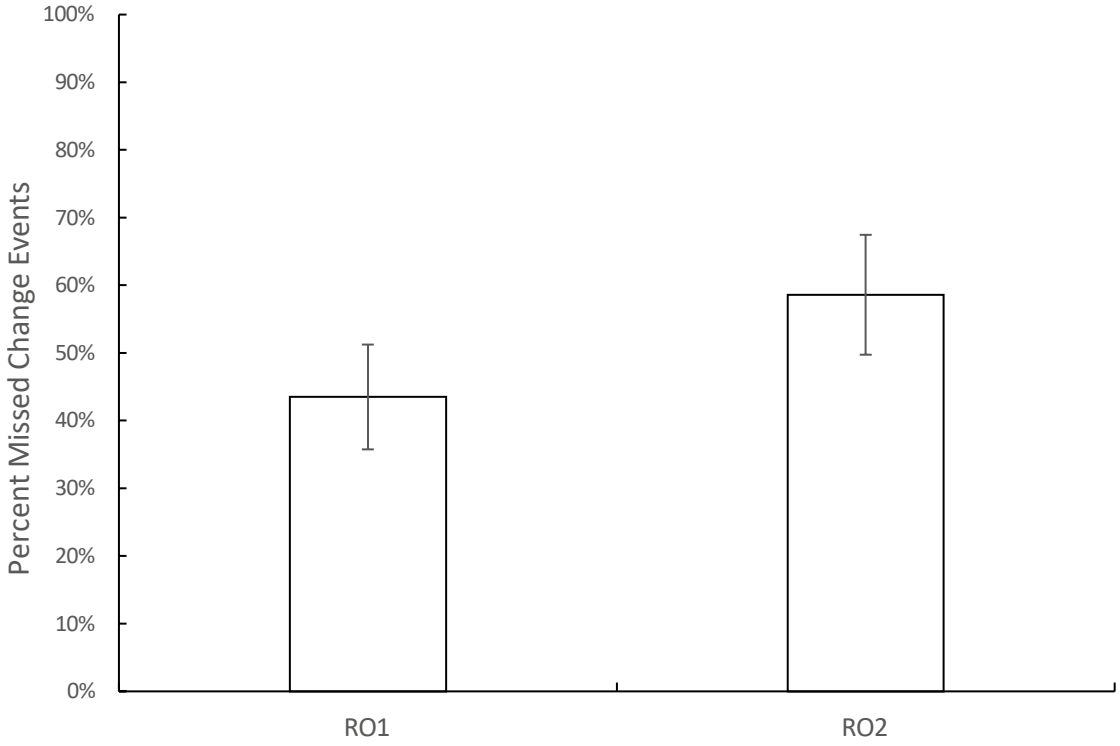


Figure 4-18 Percent missed change events by RO role (error bars denote standard errors)

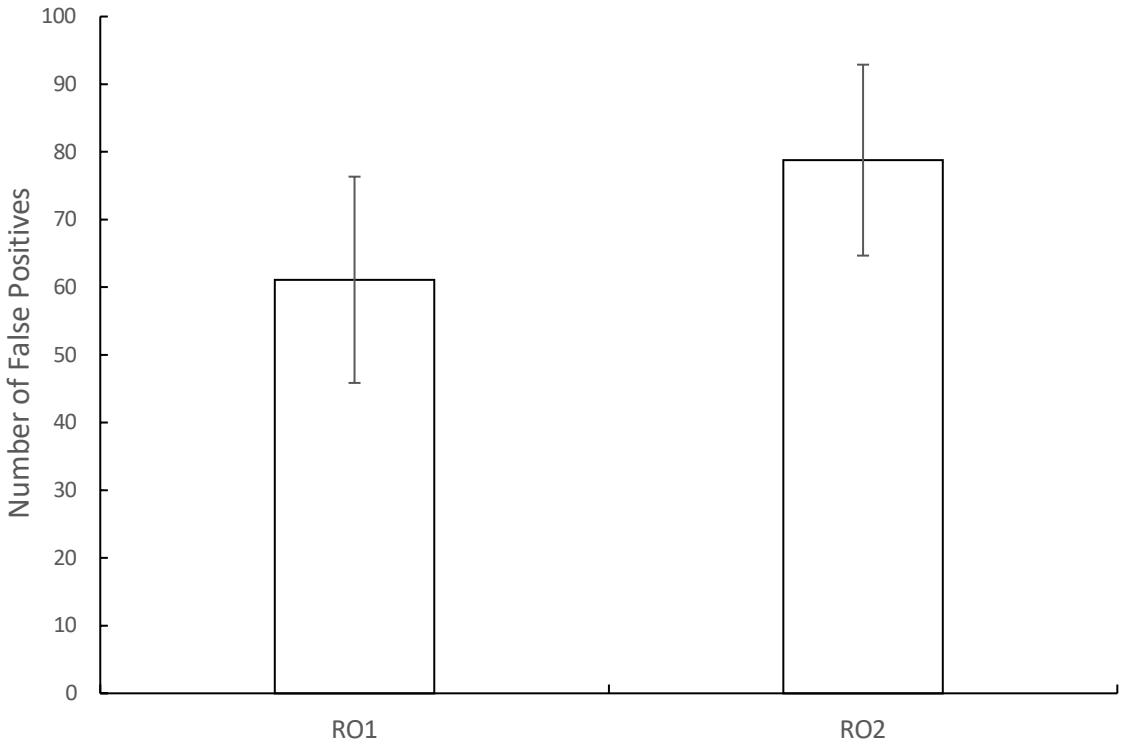


Figure 4-19 Number of false positive detections by RO role (error bars denote standard errors)

4.4.3.3.2 Response Implementation

The percent of correct manipulations, percentage of description errors, percentage of mode errors, and number of times a participant followed the correct sequence of identifying the I&C and then manipulating it in the correct direction for each participant was measured while completing the response implementation task type. No difference in correct manipulations was found for RO1 (M = 83.57, SD = 17.44) and RO2 (M = 82.22, SD = 22.79) roles (p > .05) (Figure 4-20 Percent correct manipulations by RO role (error bars denote standard errors)). No difference in description error manipulations was found for RO1 (M = 4.37, SD = 9.07) and RO2 (M = 11.11, SD = 3.70) roles (p > .05) (Figure 4-21 Mean frequency description errors by RO role (error bars denote standard errors)). No difference in mode error manipulations was found for RO1 (M = 2.22, SD = 6.67) and RO2 (M = 4.07, SD = 8.13) roles (p > .05). No difference in the number of correctly followed sequences was found for RO1 (M = 0.44, SD = 0.53) and RO2 (M = 0.56, SD = 0.43) roles (p > .05) (Figure 4-22 Percent mode errors by RO role (error bars denote standard errors)).

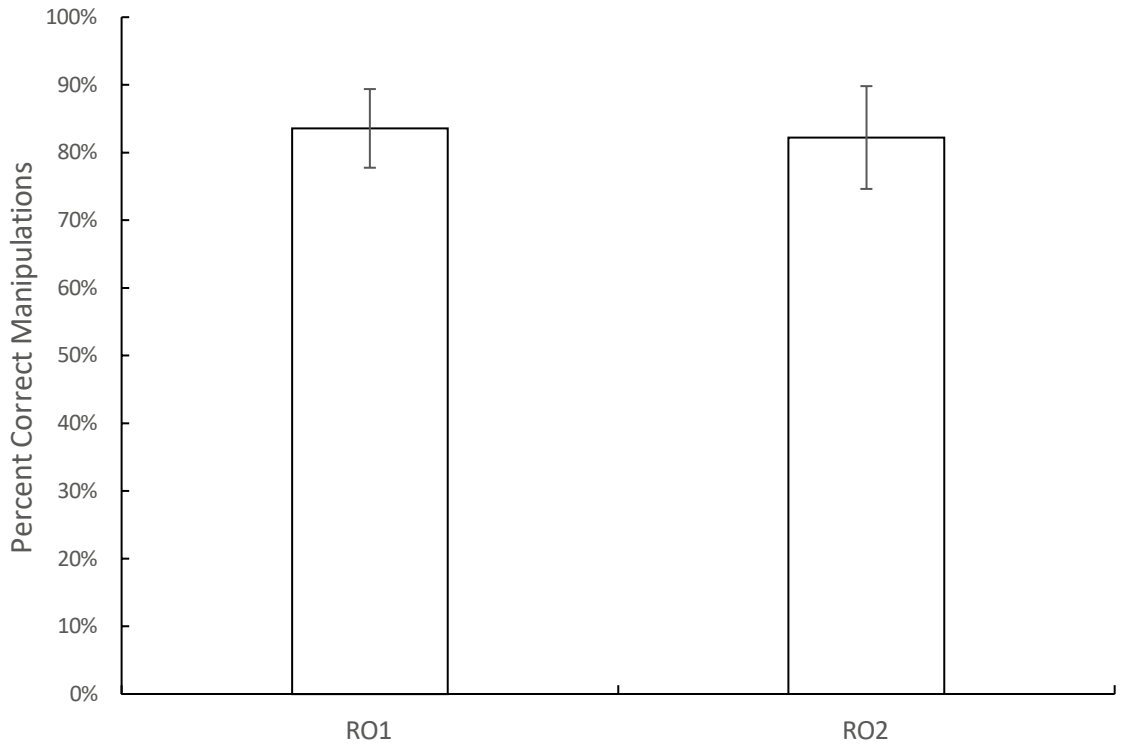


Figure 4-20 Percent correct manipulations by RO role (error bars denote standard errors)

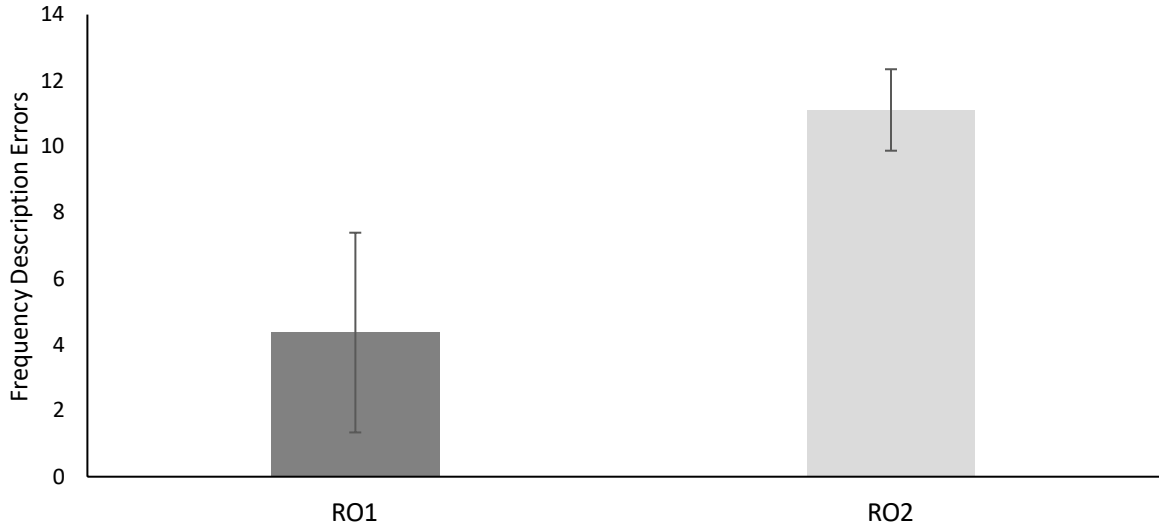


Figure 4-21 Mean frequency description errors by RO role (error bars denote standard errors)

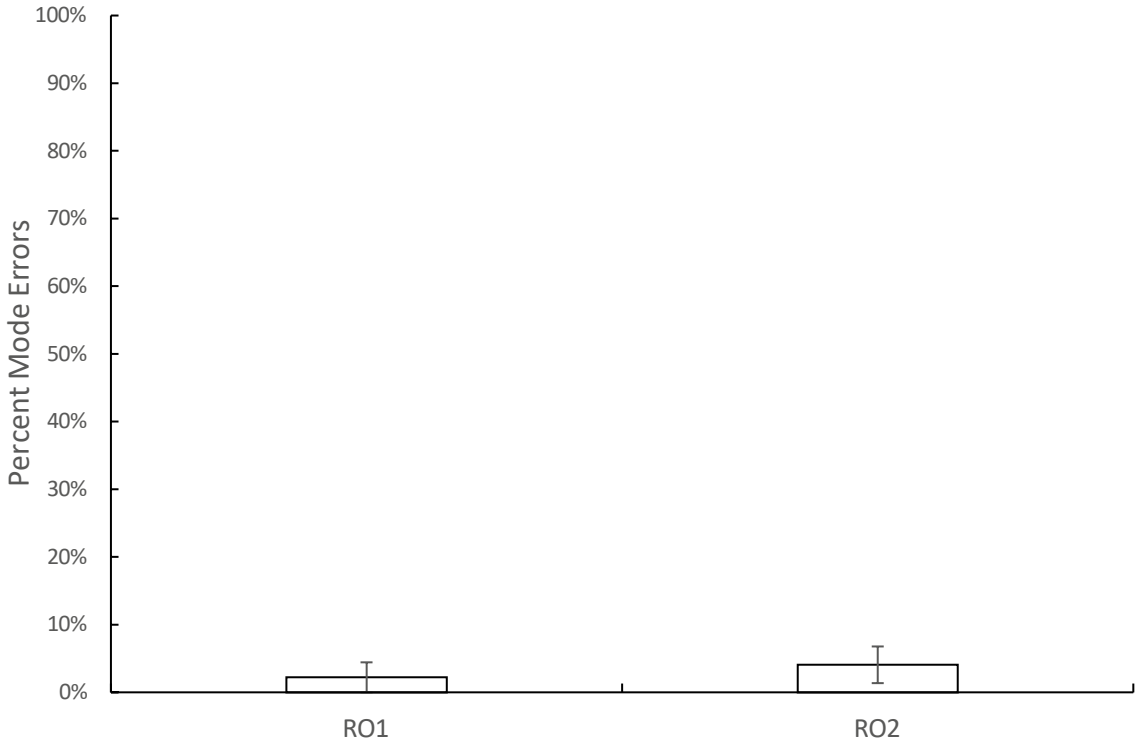


Figure 4-22 Percent mode errors by RO role (error bars denote standard errors)

4.5 Discussion

RO1 overall showed lower workload than RO2 likely due to information processing requirements being different for the checking and response implementation task types (at the step level).

Lack of consistency in valve type (spring loaded vs locked) and state for RO2 across 4 steps was likely the cause of the difference in information processing demands. These results may have implications for probing a more granular level for Human Reliability Analysis (HRA) model development. For example, when considering performance shaping factors of the role performed, it may be more important to consider the influence of the task details such as valve type and state over the particular role. Similarly, it may be important to consider the potential effect that task order may have on the information processing requirements of a particular sequence. However, task performance, navigation, and communication did not meaningfully differ between the roles overall. These trends were found across all types of measures (performance, subjective, and physiological).

4.5.1 Workload

The analysis revealed several findings for workload levels between the three task types (checking, detection, and response implementation). It was worth noting that while several measures had statistical significance between checking, detection, and response implementation task types, there was no single measure that was sufficiently effective at distinguishing between each of the three task types.

4.5.1.1 Perceived Workload

The detection task type, in general, had the highest rated workload for both RO1 and RO2 roles as indicated by the NASA-TLX and MRQ, both administered post-task. This finding of higher workload for detection is consistent with Study 1. The detection task type differed from the other two task types in several ways. First, the detection task type took much longer to complete. It lasted on average 24 minutes and 41 seconds whereas the other two task types took on average 4 minutes and 25 seconds to complete. This is because the detection included four sustained attention steps that lasted five minutes each. Within each of these four steps, the participants were required to continuously monitor a single gauge and report every discrete change in the gauge's value by pressing the gauge label directly below the gauge. The detection task had 240 discrete changes that occurred on the four gauges that needed to be reported. In addition to the sustained attention required for acknowledging gauge changes, participants were also required to keep in working memory the gauge's threshold value at which they needed to report back to the SRO once it crosses. These differences in the detection task type compared to the checking and response implementation task types are what drive the major changes in levels and types of workload.

RO2 exhibited higher workload rating on the NASA-TLX and MRQ. In particular, the MRQ indicated that the response implementation task elicited higher spatial emergent workload for RO2 than RO1. This is likely due to the valve indication of open or shut was opposite the instruction issued by the SRO for two steps of the four for RO2, but all valves were congruent with the instructions for RO1. Working experience is another potential influencing individual difference factor. In this study, for participants in the RO1 role, eight reported experience with PWRs, three had experience with BWRs, and four used to work on a carrier or submarine. For participants in the RO2 role, eight reported experience with PWRs, five had experience with BWRs, and three used to work on a carrier or submarine. In all cases those who reported experience on a carrier or submarine also reported experience on a PWR or a BWR. The majority of the sample reported experience in a PWR (N = 16) and those who reported BWR, but no PWR experience were evenly distributed across groups (see Table 4-1 and 4-2). Working experience can influence performance because operators with primarily BWR experience may have had greater cognitive interference when using the simulated PWR than the PWR operators. However, interference could also be an issue for operators with primarily PWR experience, because the generic PWR simulator was distinct in many ways compared to their former home plants. More research would be needed to fully investigate the generalizability and thus potential for interference effects of former training experiences on operator performance.

ISA ratings, which were collected by verbal response to an auditory prompt, showed no significant differences. This could be because the operators were hearing each other state their ratings and they likely were mindful of feeling judged. RO1 and RO2 were prompted in an alternating fashion and the task types were randomized per RO pair. However, they might have perceived pressure since they were supposed to be experts and thus verbally expressed less load than they might have done if answers were given completely anonymously from their peers and researchers.

4.5.1.2 Physiological Response

Overall, theta and beta in several regions of the brain showed less change from baseline for detection than both checking and response implementation. Theta and beta increases are

associated with workload (Kurimori & Kakizaki, 1995). Theta has been shown to increase with concentration and working memory demands. Beta is associated with arousal, attention, and workload. The sustained attention requirements, particularly for the detection task, is evident in the decrement in performance and physiological response (Berka, Levendowski, et al., 2007). Additionally, the detection task is longer than both checking and response implementation but has the same number of navigation steps. While the working memory component of detection is harder (as indicated by performance) than the other task types, the portion of time during the task where the participant had to hold the gauge name in working memory is less than the other task types. The working memory and concentration task demands are seen in the increase in theta across both hemispheres for the three task types. These demands are indicated by the fact that theta increased from the resting baseline. This is consistent with the fact that the three task types had the navigation component which, among other demands, also required participants to hold information in working memory. Theta has a larger increase from baseline for both checking and response implementation compared to detection. Detection's overall portion of task time for navigation was significantly less than both checking and response implementation. The larger theta increase for checking and response implementation shows that navigation task components were a major source of working memory and concentration demands.

Beta for all three task types showed an increase from baseline. This increase is a direct reflection of the cognitive processing demands imposed by the tasks. However, beta increase from baseline for detection was less than both checking and response implementation. Beta reflects the participant's arousal and attention. The fact that detection had less of an increase than both checking and response implementation, shows that the sustained attention component when monitoring a gauge is a sustained attention tasks, that appears to act like a vigilance task. Sustained attention tasks are often associated with the construct of vigilance and vigilance tasks are often associated with a drop in attention and arousal. Detection lasted around 24 minutes which shows how quickly sustained attention can deteriorate. Perceived global workload ratings were higher for detection compared to checking and response implementation, which provides further support for this task type reflecting a vigilance task.

4.5.2 Performance

Both the RO1 and RO2 subjects performed the three task types with relatively few errors. Three-way communication between task types, however, revealed that detection required a larger number of instructions to be repeated compared to checking and response implementation. This is consistent with the NASA-TLX and MRQ ratings for the detection task. Instructions for the checking task type required the participant to verify a specific valve or light box was tripped, open, or shut. Instructions for the response implementation task type required the participant to verify then open or shut a specific valve. Instructions for the detection task type required the participant to verify then report back once a specific gauge crosses a threshold value. The difference for the detection task type instruction was that participants were required to remember two numbers, which included a value of the threshold to monitor and report back in addition to the number in the gauge name. It was observed that the repeat back of the initial SRO instruction would often fail at the threshold value before a participant would request a repeat instruction (e.g., "SRO, understood you want me to verify LI 494 Sierra Alpha and report when…SRO, Pease repeat"). While not part of experimental training, ROs would often request just the report back threshold during the detection task (e.g., "SRO please repeat the threshold for LI 494 Sierra Alpha").

4.6 Conclusions

Similar to Study 1, the detection task type had the highest rated workload regardless of operator role. It should be noted that the detection task required the longest time to complete. It was operationally distinct from the other two task types in that participants were required to keep the threshold value (the number on the gauge that required RO action) in working memory. The cognitive requirements, because of the higher working memory demand associated with detection task performance were greater in comparison to checking and response implementation.

In terms of physiological response, theta power had a larger increase from baseline for both checking and response implementation compared to detection. Theta increases are associated with concentration and working memory demands. Navigation and identification task components were the major source of working memory and concentration demands. Future work might evaluate the methods to reduce navigation and detection to determine if they yield the anticipated benefits. As a reflection of the participant's arousal and attention, beta increase from baseline for detection was less than for both checking and response implementation. The sustained attention component when monitoring a gauge during the detection task acts like a vigilance task.

Overall, the former operators (who consisted of formerly licensed NPP and navy/nuclear operators) in Study 2 and the experienced participants from Study 1 showed similar trends in the performance and workload data.

The findings reported here have applicability for modernization of current and new NPP control room designs and future research using digital simulators. This research found experts relied on their past training to prioritize tasking such that critical tasks are completed with more diligence than other less critical tasks. This is seen in the fact that almost all the expert participants identified the correct control on their first attempt. While experts were diligent in their critical control room tasks, several expert participants consistently omitted the recipient with their communications. This could be evidence of a more refined approach to task prioritization (relative to novices) or an artifact of the contrived experimental environment.

4.6.1 Strategies

The researchers observed that several participants practiced different strategies while performing the experimental tasks. These strategies appear to be learned behavior from participants' years of training at their home plant but are not regulated behaviors or industry standards. The first strategy was to hover their finger near the I&C they were instructed to locate. Then they would complete the verbal communication with the SRO before completing the identification action, likely as a way to verify they had identified the correct I&C. Only once the SRO said "RO that is correct" would the RO touch I&C to signal they located it. A different but related strategy that was observed was requesting a repeat of an instruction from the SRO once they found the desired I&C. This practice was observed mostly with I&C that took a considerable amount of time to locate. These strategies show that both relevant training strategies and NPP knowledge assist in rule-based tasks. These two strategies indicate that ROs value identifying the correct I&C even at the cost of redundant information being communicated.

During the detection task, participants were required to monitor gauge changes and report once a gauge fell below a threshold. The gauge changes were scripted for repeatability between participants but were based on timing from the full-scope physics-based reality of the would-be NPP simulator. The simulator was run with the same initial conditions and followed the same procedures as defined in the experiment. Gauge values would jump up and down in small discrete increments but trend downward. This is because the gauge changes were based on plant physics. Each gauge script concluded exactly at five minutes. The last gauge change crossed the reporting threshold but was relatively close to the value. A strategy some participants would employ during the detection task was to wait for 30 to 45 seconds after the threshold was crossed before reporting back to the SRO. Participants were waiting a reasonable amount of time to make sure the gauge value was stable and below the threshold.

The last strategy that was observed was periodic reporting back to the SRO intermediate updates on the gauge value and its trending direction every minute. The use of these last two strategies were linked and one possible explanation for the variability in use of these particular strategies was explained by operator experience or recency of experience. In other words, the more experienced operators or those who had more recently operated a plant, were more likely employing the use of these strategies. Further research could be conducted to delve more deeply into this phenomenon.

5 GENERAL DISCUSSION AND CONCLUSIONS

5.1 General Discussion

The NRC procured two identical GSE GPWR simulators, one of which is housed at the NRC and the other at the University of Central Florida as a part of the HPTF contract. This project is ongoing and will extend the findings reported in the current document. The present report focused on the outcome of two large-scale experiments that were conducted to address challenges associated with developing a research methodology for using novices in a highly complex, expert driven domain. To do this, we focused on rule-based and skill-based tasks instead of those requiring domain knowledge and experience. The three tasks were checking, response implementation, and detection. The exact I&Cs for the scenario were determined via a task analysis and cognitive task analysis. The operating procedure selected was chosen using expert elicitation with SMEs and because it required the fewest number of panels, and could easily be scaled up or down, depending on the participant population. The initiating condition was a loss of all alternating current power. Levels and types of workload were of primary interest as this is one factor evaluated in the regulatory verification and validation guidelines in NUREG-0711 and high workload is usually associated with poor performance and errors. Workload was measured subjectively with questionnaires and objectively with performance and physiological measures. Across all the experimental sessions, regardless of interface type or operator role, the detection task was always the most difficult of the three tasks.

The detailed results suggest that the number and order of task type impact workload and task performance, thus increasing operator vulnerability to error. These insights have the potential to inform the quantifiable approach of human error probabilities used in HRA. For example, multiple detection tasks should not be placed near each other, as the error likelihood may increase (i.e., compound) over time. More practically speaking, when distributing tasks among operators within a procedure or scenario, consideration should be given to ensure that the detection tasks be distributed among available operators. Similarly, checking tasks should also be separated in time or distributed across available ROs because the immediate response required elicits higher temporal demand and could mean that errors are more likely to occur.

Studies 1 and 2 also provided an initial examination of workstation design (e.g., sit-down versus stand-up) and interaction control techniques (e.g., mouse cursor and click versus touchscreen) as might be associated with new soft control designs. See Section 2.4.1 ‘Defining the NPP Simulated Environment’ for a description of the characteristics that define NPP simulated environment. The workload and ergonomic impacts of different workstation and input design is important for future facing regulatory guidance. A sit-down desktop workstation with mouse and keyboard for inputs (i.e., interaction control) was evaluated compared to large touch screens for input with stand-up workstation. Accuracy was better with the desktop, but the workload was lower in the touchscreen as reflected by subjective and physiological measures. These findings carry implications for advanced reactor designs which bring with them new concepts of operation accompanied by new technology and interfaces. Future research can further investigate the impact of input design and workstation ergonomics on operator performance, perceived, and physiological workload.

The study findings also highlight some general issues in workload assessment and methodology in the NPP context. First, multiple methods for assessment were utilized. Second, the current research has secured data from both novices and more experienced populations. Third, the current study identified some differences in workload response between RO1 and

RO2 operator roles. How effective is the current workload strategy for identifying operationally significant differences between operator roles? These questions will be addressed in relation to both Studies 1 and 2 of this series. Each of these issues is discussed in the context of how they can support NRC's regulatory mission in the subsequent sections.

5.1.1 Multifactorial workload assessment for plant operations

As discussed in 2.3.3, self-report scales, especially the NASA-TLX (Hart & Staveland, 1988), are the most popular workload measures, for their ease of use and demonstrated capacity to identify overload situations (Estes, 2015). Adding physiological measures adds to costs, and thus requires justification. Workload measures can be evaluated against multiple standards described by Eggemeier, Wilson, Kramer, and Damos (1991). The most relevant of these in the present context are sensitivity, diagnosticity, and selectivity.

5.1.1.1 Sensitivity

Sensitivity refers to the extent to which the workload measure registers true differences in task demands. In the current study, multiple subjective and objective measures identified the higher workload associated with the detection task, but instrument sensitivities differed, as measured by the effect size statistics (η_p^2). Generally, a larger effect size indicates greater sensitivity. The effect size demonstrated by the NASA-TLX was substantial ($\eta_p^2 = .35$). However, other measures showed substantially larger effects, including fNIRs. For the right hemisphere response, the effect size was .57. Conversely, the ISA, which has demonstrated sensitivity in some contexts (Tattersall & Foord, 1996) failed to discriminate the task conditions at all.

Among the three subjective measures, the NASA-TLX global workload and short-term memory process subscale in MRQ showed a significant difference. The NASA-TLX tended to be sensitive to task demand variations in NPP operation domain. This was consistent with Ikuma, Harvey, Taylor, and Handal's (2014) finding that the NASA-TLX was sensitive to task demand changes. The operator's subjective experience of "busyness" may not reflect the level of frontal-cortical activity indexed by fNIRS. Similarly, not all physiological measures were sensitive to the reported manipulations. A case in point is CBFV which in other contexts is sensitive to loss of vigilance during sustained attention tasks (Warm et al., 2012). The insensitivity of fNIRS in the current study could indicate that the threat to vigilance posed by cognitive fatigue and attentional resource depletion is not a major one, even on the detection task, probably because the task was of relatively short duration. Traditional vigilance tasks unfold over hours and hundreds or thousands of trials, follow up research related to vigilance specifically would be needed to determine if the insensitivity of fNIRS is related to task parameters or the role of fatigue and attentional resource depletion during the tasks tested. The practical significance is that choice of a high-sensitivity measure such as fNIRS is essential for investigations of how detection workload might be mitigated, or how workload might change as new technologies such as automation are introduced.

Similar differences in sensitivity across measures were obtained in Study 1. For example, the touchscreen condition analyses found that fNIRS and some EEG measures showed greater sensitivity to task type differences than did the NASA-TLX. Experiment 2 also showed sensitivity differences in detecting the effects of interface type on workload, i.e., desktop vs. touchscreen. In this case, the analysis showed only a modest effect of interface on NASA-TLX scores, with higher workload for the desktop. Again, substantially larger effects were found for fNIRS, suggesting that this method is especially suitable for interface evaluation.

HRV was sensitive to variations in workload in Study 1 but not Study 2. There are several potential reasons for this difference between the two studies. It could be related to differences in NPP experiences between the novice and expert study groups, such that participants in Study 1 produced larger HRV deviations as a function of condition because of their inexperience in the domain. The experts in Study 2 may have produced less dramatic condition-related variations that was not detectable through this physiological measure.

5.1.1.2 Diagnosticity

Diagnosticity is defined as the extent to which the workload measure identifies the source of workload. Global measures such as those provided by the NASA-TLX (Hart & Staveland, 1988) and the ISA (Tattersall & Foord, 1996) reflect a subjective integration reflecting multiple demands on the operator. Thus, their diagnosticity is limited; major sources of overall workload may differ in different contexts (Eggemeier et al., 1991). Indeed, changes in NASA-TLX workload may be accompanied by quite different patterns of change in physiological response, implying that it is not very informative about neural response to task demands (Matthews et al., 2015). This makes sense, given that the NASA-TLX ratings are based on a self-assessment of perceived resources demanded by the task. The perceived and actual neurophysiological demands would not necessarily be correlated as they are different aspects of the workload construct. Despite this, the NASA-TLX provides some level of diagnosticity through its six subscales. The analysis identified mental demand and frustration as the principal sources of workload, but task type effects were similar across all six scales; the NASA-TLX did not uniquely identify the source of higher demand for the detection task.

The study illustrates the benefits of complementing the NASA-TLX with the MRQ. The latter instrument showed that demands of the detection task were especially high for the spatial concentrative and spatial quantitative subscales. Operationally, this finding implies that efforts to improve the design of the interface should focus on its spatial aspects. The organization of information is always important, but in large format information systems, such as control rooms the organization, scale, prominence, and persistence (hierarchical vs. persistent displays) all impact the extent of the spatial processing demands. Additionally, these kinds of large format information systems, regardless of digital or analog state require significant time to navigate, which increases the baseline memory and attentional demands and necessitates efficiency improving navigation elements to reduce this load (see Radle, Jetter, Butscher, Reiterer, 2013). Similarly, understanding the neurocognitive bases of the different physiological responses enhances diagnosticity. As discussed in earlier sections, the pattern of EEG response to the three different tasks identifies the challenge of sustaining attention as one of the factors potentially contributing to workload for the detection task. The enhanced fNIRS response to detection may similarly represent demands on sustained attention.

Similarly, analysis of the pattern of workload change across multiple indicators demonstrates diagnosticity in other studies (e.g., Matthews, Reinerman-Jones, Wohleber, Lin, Mercado, Abich (2015). In Study 1, for both the touchscreen and desktop conditions, workload differences were especially evident in NASA-TLX performance and effort ratings (i.e., diagnostic to the source of WL), in MRQ short-term memory and spatial scales, and in EEG and fNIRS response. In this case, though, there was a contrast between subjective and physiological measures. NASA-TLX and MRQ data suggested that participants experienced higher demands with the desktop interface, but the EEG and fNIRS analyses implied higher brain activity while using the touchscreen. Performance tended to be better on the desktop than the touchscreen (depending

on the measure) so that the elevated subjective (and to some extent physiological) workload may be associated with effective task-directed effort. By contrast, the increases in brain activity associated with the use of the touchscreen could reflect the greater physical demand of using a touchscreen, the novelty of engaging with control room elements in that way, or many other factors in addition to participant and display characteristics.

5.1.1.3 Selectivity

Selectivity is the extent to which the measure represents a unique workload response without also reflecting related factors such as stress and physical activity. In the present experimental series, selectivity proved to be a lesser concern, as the work environments were not especially challenging or unpleasant, limiting stress response. Workload levels were generally low to moderate so that participants were unlikely to feel overwhelmed or upset by poor performance.

ECG HR data illustrate this issue. This measure is vulnerable to a lack of selectivity. The sympathetic activation associated with stress elevates HR. So too does physical activity.

The current study failed to demonstrate effects of the independent variables on HR implying that (1) HR is insensitive to workload in this context, (2) task manipulations were not especially stressful, and (3) tasks were not associated with gross metabolic differences driven by physical activity. In other contexts, HR might be more sensitive to operational factors; for example, an HR response disproportionate to other workload indices might signal excessive stress.

Overall, the findings reinforce the message that single workload indices rarely provide an adequate picture of operator response to task demands, especially when workload levels are moderate, and operational issues are more complex than simple overload. The NASA-TLX is effective in identifying situations where the operator is overwhelmed by task demands but less useful for the detailed identification of different workload sources. The MRQ is especially diagnostic in identifying the sources of cognitive demands, and physiological measures are diagnostic for neurocognitive response. Careful analysis of the pattern of results, and triangulation of workload responses with performance effects, may provide the optimal strategy for evaluating control room design and interventions to mitigate workload and enhance performance. Establishing consistency of results also contributes to the aim of ensuring that workload experimentation can be applied to a range of plant types, designs, and indicators.

5.1.2 Utilization of novice samples in the assessment of workload issues

As discussed previously, an overarching goal for this research is to explore the feasibility of using novice participants to assess the workload associated with common, skill-based, NPP operator tasks (excluding knowledge-based tasks). A key advantage of doing so is to facilitate comprehensive and systematic investigation of possible emerging workload issues associated with control room modernization, is that it is impractical to run extended series of studies with adequately large samples of trained operators (see discussion in section 1.2.3). For example, the large sample of Study 1 was necessary to achieve sufficient statistical power to provide a comprehensive evaluation of the impact of interface type on workload.

A comparison of findings across the present set of experiments validates drawing inferences about workload from samples. An obvious concern is that workload will simply be much higher for novices as a result of their lack of experience and practice, relative to experienced operators so that no generalization to operational settings is possible (Matthews et al., 2019). However,

workload levels were low to moderate for all samples. In Study 1, the samples showed mean overall NASA-TLX workload levels in the 30-40 range, depending on task conditions. For the more experienced sample here, the mean for detection was similar (37.5) but somewhat lower for checking (23.7) and response implementation (25.0). A decrease in workload with expertise might be anticipated, but the means are not so different as to present a threat to generalization. Both novices and actual operators experience the task as being fairly undemanding. Similarly, all samples achieved reasonably good levels of proficiency in task performance.

Some caution is advised in making direct quantitative comparisons of workload metrics across samples of different compositions. Because of the limited sample size of experienced operators, the standard errors of means (inaccuracy of measurement) are relatively high and estimates are imprecise, compared to the larger novice samples. In addition, because of the small sample size the aggregate scores may be disproportionately affected by the responses of individual operators. Another caution with the operator data is that their previous plant experience could create cognitive interference, where interactions in the experimentally configured control room feel more effortful because it is different from the control rooms in which they have previously worked.

What may be more important than exact quantitative similarities is the reproduction of qualitative differences between task conditions in different samples. This criterion is satisfied by the comparability of task type effects in novices (Study 1) versus more experienced operators (Study 2). In both samples, evidence was found from multiple workload measures for the greater demands of detection, compared with checking and response implementation tasks.

The feasibility of utilizing novice samples depends on the principles elaborated in Study 1. That is, investigations are limited to procedural tasks that can be readily trained. Quality and provision of training must be sufficient, and criteria for adequate performance in training must be applied to weed out the minority of participants unable to acquire the needed skills. The simplification of control panels described in a previous technical report (Reinerman-Jones & Mercado, 2014) is also important.

5.1.3 Influence of operator role

A novel feature of the Study 2 was the opportunity to compare workload responses in operators performing in RO1 and RO2 roles. (Study 1 assigned the RO2 role to a trained confederate). RO1 and RO2 performed different steps of the simulated scenario as detailed in 2.8.1.1, but these steps were similar. Additionally, RO1 and RO2 performed the same tasks, thus, substantial differences in workload were not anticipated.

As expected, there were no major workload differences associated with operator role, however, there were subtle differences. The reported MRQ data in section 4.4.1.1.3 demonstrated workload differences for the response implementation task that could be attributed to incidental differences in whether valve settings were congruent with the SRO's instructions to open or shut the valve. The overall NASA-TLX workload was also higher for RO2s than for RO1s. In fact, the mean for RO2s (37.9) was similar to the novice sample in Study 1, whereas the means for RO1s (19.6) was substantially lower. Generally, these comparisons reinforce the caution made above, that too much weight should not be placed on comparisons of means based on small samples. However, the lower mean for RO1s also points out the potential hazards of underload in the operational setting. Low workload may be associated with boredom and distractibility and

can threaten operator performance if sustained over extended time periods (Hancock & Warm, 1989).

We also compared novice to experienced (student employees of UCF who were extensively trained and completed 30 sessions by serving as the study confederates in Study 1) to expert (former operators employed by the NRC) operators Experienced operators performed the best and had the lowest workload, but expert data trended in the same direction. The explanation for this finding is that for all but one expert, this was not their “home” plant, whereas the experienced participants had only ever performed these tasks on this specific simulator, making them better performers than the experts in this one narrow context. What this training effect and the success of the experienced participants demonstrates is that for the purposed of rule-based and skill-based tasks, novice and experienced participants are viable. The individual who came from the plant upon which the simulator was based did perform equivalently to the experienced participants. Furthermore, operators who had formerly operated technology other than PWR NPPs (e.g., BWRs or navy nuclear submarine reactors) had the most difficulty in learning the scenario. In particular, errors were made on color and switch direction for reporting open or closed. This indicated that training is especially important to reduce the negative transfer of previous learning to a new interface and system.

5.1.4 Further implications: Human reliability analysis

The primary focus of the current effort has been on the methodological issues for workload assessment discussed in the preceding sections, and on specific demand factors of task type and interface. However, there are also wider implications for the NRC mission in evaluating future power plant designs as technology advances. As the information in the MCR becomes increasingly integrated, new designs and technology are introducing new concepts of operations. Prospective innovations, especially increasingly intelligent automated systems, may raise new challenges for human factors evaluations (e.g., Matthews et al., 2016). Transformative technological change requires more than piecemeal assessments of specific demand factors, such as interface type.

In the NPP context, Tran et al. (2007) highlight the potential of HRA as an approach to assessing and minimizing risk in next generation control rooms. The original focus of HRA was on probabilistic human error-rate prediction, but contemporary approaches aim to model the cognitive processes that underlie human behavior, and so must incorporate factors that may impair processing including workload and stress (Mosleh & Chang, 2004; Whaley et al., 2016; Xing et al., 2020). Compatible with the current methodology, Tran et al. (2007) emphasize the utility of physiological measures of workload in model development. These authors point out that factors influencing performance are often dynamic and interdependent, and the continuous monitoring of operator state afforded by psychophysiology provides a means for tracking factors dynamically. An example here is that demands on the operator may be influenced by task sequencing. In the current experimentation, workload elevation in the detection task may, in part, reflect the repetition of the monitoring assignment across repeated 5-min blocks.

The current research also highlights the challenge of assessing operators functioning at different levels of granularity. Some workload research adopts the coarse-grained technique of assessing overall workload only, e.g., with the NASA-TLX (Hart & Staveland, 1988). As discussed in 5.1, this method is demonstrated to be inadequate for the NPP domain. The more granular multivariate approach advocated here is an improvement, and temporal modeling of demand factors as advocated by Tran et al. (2007) may represent a further enhancement. For

example, dynamic modeling may be especially important for interacting with systems able (within certain bounds) to make autonomous decisions and communicate them to the operators (Matthews et al., 2016). However, the present research also indicates how, in some contexts, differentiating multiple workload components may be insufficiently granular. For example, it has been discussed how the finding of lower workload for RO1s than for RO2s is likely due to process requirements being different for checking and response implementation task types, at the step level.

As another example, the current study found differences in the strategies used by experts to perform tasks (4.6.1). Strategies included hovering the finger near the relevant I & C waiting for verbal confirmation prior to identification. Some experts may also wait to report a below threshold value during detection to see if the gauge readout increases. These observations indicate the importance of including variation in the knowledge base that experts apply to rule-based tasks in modeling the impact of demand factors and the likelihood of error. Of course, studies of novices cannot contribute here. Alternatively, research might identify the most effective strategy and recommend that it is trained universally.

However, in considering HRA methodologies, a few results should be noted. Findings across both Studies 1 and 2 indicate that:

- Regardless of interface, role, or level of experience, the detection task type was the most demanding across measures.
- The checking task type elicited the greatest temporal demand across measures. This information can inform HRA models for procedure design and evaluation because a procedure that has sequential detection (e.g., 3 detection tasks back-to-back) or checking (e.g., 5 checking tasks back-to-back) steps will likely yield greater error potential.
- Detection and checking steps issued overlapping instructions to RO1 and RO2 (e.g., SRO instructs RO 1 detect when gauge XYZ reaches N and then immediately instructs RO2 detect when gauge ABC reaches N) might prevent effective teamwork behaviors that mitigate errors.

Further research is needed to understand the maximum number of sequential same task type steps that should be permitted in a procedure and to determine if there is a maximum number of steps of a specific task type that should be allowed in a single procedure. Additionally, the order in which task type steps are executed might impact error rates. For example, does checking, response implementation, detection, checking, detection, response implementation elicit greater workload, and thus increased error than detection, checking, response implementation, checking, response implementation, detection? The outcomes of those investigations would inform the HRA model or algorithm development for procedures but would also inform I&C design and NPP functions.

5.2 General Conclusions

5.2.1 The “Workload Picture” and the Measures

This section presents some take-aways with regards to the “workload picture” in terms of developing assessment methods that are robust across different task types, interfaces, and samples. As a general conclusion, selection of individual measures should be based on the overall assessment goals and practical situational constraints. For example, fNIRS might be the most sensitive technique for a given situation, however, this type of physiological technique is

burdensome for the participant and produces data that can be difficult to interpret. The assessment strategy has to balance the need for sensitivity with these types of practical considerations.

It is important to understand the nature of the demands imposed by specific tasks. For example, the detection task might be considered similar to a vigilance task. One of the physiological measures we chose, fNIRS measures the oxygenation in the prefrontal cortex. The prefrontal cortex is used for management of attentional resources and executive control of cognition. The measure that was sensitive to prefrontal cortex activities was sensitive to changes for the detection task versus the other two tasks, demonstrating how physiological measures can be used to measure workload changes associated with different task types.

The subjective NASA-TLX measure also found a main effect for the detection task for the global workload. The NASA-TLX, especially the frustration subscale, captured the perceived workload change induced by detections, which may suggest that NASA-TLX is sensitive in detecting workload response in specific types of tasks. Frustration is a particular hallmark of a vigilance task (Warm, Dember & Hancock, 1996; Warm, Parasuraman, and Mathews, 2008). These results further support the notion that the detection task is a type of vigilance task involving sustained attention and can be measured using these techniques.

For the comparison of interface types, the NASA-TLX indicated that global WL was higher for the desktop interface. The physiological measure of HR also found greater increases in HR from baselines for the desktop when compared to the touchscreen. Future investigations may consider the underlying reason for this observation. Overall, the findings reinforce the message that single workload indices rarely provide an adequate picture of operator response to task demands, especially when workload levels are moderate, and operational issues are more complex than simple overload. Future research in more realistic operational environments should be conducted to determine if these conclusions are generalizable. There may be important implications for workload measurement in the context of Integrated System Validation for new designs and future plants.

Finally, in terms of workload there was consistency across the findings from both studies 1 and 2. This means that the suite of workload assessments applied in this experimental context would likely be appropriate for use across a wide range of plant types, designs, and indicators.

5.2.2 Future Directions: Refinement of Workload Assessment Methodology

The long-term, overarching objective for the HPTF is to support safe plant operations by examining challenges related to the impact of technology upgrades, automation of tasks, and digital interfaces on human operators' ability to perform monitoring and control functions in the MCR. The near-term next steps for this program might be to explore the generalizability of the results from the previous studies in an environment closer to the reality of true NPP operations. More specifically, studies may further validate the NPP simulator and methodology used and generalize the findings from the full-scope, reduced size (i.e., simulated analog) simulator with a hierarchical I&C layout to the full-scope, full-size, analog hard panel simulator with a spatially dedicated and continuously visible I&C layout. To do so, it will be necessary to conduct a follow-on series of experiments to systematically compare the results obtained from Studies 1 and 2, which used novices and former operators operating in a simulated analog interface, to a full-size simulator consisting of traditional bench boards with hard-wired I&C with former operator participants.

As the initial objective was, in part, initially to establish a baseline, simplification was needed in order to be able to measure with certainty. For example, experimental control in the form of task blocking was used such that direct measurement could be done for each task type. So, the next steps might be to use similar measures, but in an environment more representative of real NPP operations. Specifically, participants would perform the same task types by stepping through a full procedure/scenario without task blocking in a full-scope, full-size, analog simulator. The methodology may be further validated if similar trends are observed for the measures and objective performance results (e.g., Detection Task type most demanding) but in a more realistic operating environment both in terms of the simulator interface and also because they will be performing a full scenario i.e., no blocking of task types. Furthermore, we may determine if the measures of workload are similarly sensitive to the task types and the operating environment in the same way (e.g., NASA-TLX and fNIRS were the most sensitive overall regardless of task type in the first studies; EEG and fNIRS were most sensitive to the detection task as a marker of the need for sustained attention).

5.2.3 Methodological Conclusions

Nuclear specific human performance data collection efforts large enough for quantitative analysis are not widely practiced. The staff at the NRC determined it necessary to develop its own such research program with the hope that others might follow suit. Our focus was to develop a methodology to gather meaningful data from novices using a simplified operating environment to inform us about the highly complex operational environment of the NPP MCR.

Using this research design strategy to develop a baseline, we anticipate being able to identify measures of workload best suited for particular tasks or a combination of tasks, the levels of workload associated with tasks, and the kind of workload induced (e.g., physical, cognitive) by tasks. Further, we expect that our method will improve data collection techniques for use with the operator population, such that lab results may be further validated.

The methodology presented in this RIL can serve as a foundation for future human factors testing in the NPP domain and other domains that involve complex systems and team operations. This work will expand the understanding of performance in complex systems operations and explain how factors such as new technology or concepts of operation impact performance.

6 REFERENCES

Abich IV., J. (2013). *Investigating the universality and comprehensive ability of measures to assess the state of workload* [Doctoral dissertation, University of Central Florida]. https://stars.library.ucf.edu/etd/3004

Aldrich, T. B., Szabo, S. M., & Bierbaum, C. R. (1989). The development and application of models to predict operator workload during system design. In G. R. McMillan, D. Beevis, E. Salas, M. H. Strub, R. Sutton, & L. Van Breda (Eds.), *Applications of Human Performance Models to System Design* (pp. 65–80). Springer US. https://doi.org/10.1007/978-1-4757-9244-7_5

Ayaz, H., Shewokis, P. A., Curtin, A., Izzetoglu, M., Izzetoglu, K., & Onaral, B. (2011). Using MazeSuite and functional near infrared spectroscopy to study learning in spatial navigation. *Journal of Visualized Experiments*, *56*, 3443. https://doi.org/10.3791/3443

Ayaz, H., Willems, B., Bunce, S., Shewokis, P. A., Izzetoglu, K., Hah, S., Deshmukh, A. R., & Onaral, B. (2010). Cognitive workload assessment of air traffic controllers using optical brain imaging sensors. In T. Marek, W. Karwowski, & V. Rice (Eds.), *Advances in understanding human performance: Neuroergonomics, human factors design, and special populations* (pp. 21–32). CRC Press.

Barbé, J., Chatrenet, N., Mollard, R., Wolff, M., & Bérard, P. (2012). Physical ergonomics approach for touch screen interaction in an aircraft cockpit. *Proceedings of the 2012 Conference on Ergonomie et Interaction Homme-Machine*, 9–16. https://doi.org/10.1145/2652574.2653402

Berka, C., Davis, G., Johnson, R., Levendowski, D. J., Whitmoyer, M., Fatch, R., Ensign, W., Yanagi, M. A., & Olmstead, R. (2007). Psychophysiological profiles of sleep deprivation and stress during marine corps training. *Sleep*, *30*(A132).

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., Olmstead, R. E., Tremoulet, P. D., & Craven, P. L. (2007). EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks. *Aviation, Space, and Environmental Medicine*, *78*(5, Supplement), B231–B244.

Billings, C. E., & Cheaney, E. S. (1981). The information transfer problem: Summary and comments. In C. E. Billings & E. S. Cheaney (Eds.), *Information transfer problem in the aviation system* (pp. 85–94). National Aeronautics and Space Administration.

Boles, D. B. (1991). Factor analysis and the cerebral hemispheres: Pilot study and parietal functions. *Neuropsychologia*, *29*(1), 59–91. https://doi.org/10.1016/0028-3932(91)90094-O

Boles, D. B. (1992). Factor analysis and the cerebral hemispheres: Temporal, occipital and frontal functions. *Neuropsychologia*, *30*(11), 963–988. https://doi.org/10.1016/0028-3932(92)90049-R

Boles, D. B. (1996). Factor analysis and the cerebral hemispheres: “Unlocalized” functions. *Neuropsychologia*, *34*(7), 723–736. https://doi.org/10.1016/0028-3932(95)00136-0

Boles, D. B. (2002). Lateralized spatial processes and their lexical implications. *Neuropsychologia*, *40*(12), 2125–2135. https://doi.org/10.1016/S0028-3932(02)00051-9

Boles, D. B., & Adair, L. P. (2001). The multiple resources questionnaire (MRQ). *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *45*(25), 1790–1794. https://doi.org/10.1177/154193120104502507

Borghini, G., Vecchiato, G., Toppi, J., Astolfi, L., Maglione, A., Isabella, R., Caltagirone, C., Kong, W., Wei, D., Zhou, Z., Polidori, L., Vitiello, S., & Babiloni, F. (2012). Assessment of mental fatigue during car driving by using high resolution EEG activity and neurophysiologic indices. *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 6442–6445. https://doi.org/10.1109/EMBC.2012.6347469

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology*, *42*(3), 361–377. https://doi.org/10.1016/0301-0511(95)05167-8

Burgy, D., Lempges, C., Miller, A., Schroeder, L., Van Cott, H., & Paramore, B. (1983). *Task analysis of nuclear-power-plant control-room crews: Project approach methodology* (NUREG/CR-3371). U.S. Nuclear Regulatory Commission.

Cain, B. (2007). *A review of the mental workload literature* (No. ADA474193). Defence Research and Development Canada Toronto Human System Integration Section.

Causse, M., Chua, Z., Peysakhovich, V., Del Campo, N., & Matton, N. (2017). Mental workload and neural efficiency quantified in the prefrontal cortex using fNIRS. *Scientific reports, 7*(1), 1-15. https://apps.dtic.mil/dtic/tr/fulltext/u2/a474193.pdf

Chance, B., Zhuang, Z., UnAh, C., Alter, C., & Lipton, L. (1993). Cognition-activated low-frequency modulation of light absorption in human brain. *Proceedings of the National Academy of Sciences*, *90*(8), 3770–3774. https://doi.org/10.1073/pnas.90.8.3770

Chourasia, A. O., Wiegmann, D. A., Chen, K. B., Irwin, C. B., & Sesto, M. E. (2013). Effect of sitting or standing on touch screen performance and touch characteristics. *Human Factors*, *55*(4), 789–802. https://doi.org/10.1177/0018720812470843

Derouin, A., & Salway, A. (2018). Enhancing workload assessments for validation activities associated with DBA and BDBA scenarios. *Nuclear Technology*, *201*(2), 165–173. https://doi.org/10.1080/00295450.2017.1413922

Eggemeier, F. T., Wilson, G. F., Kramer, A. F., & Damos, D. L. (1991). Workload assessment in multi-task environments. In D. L. Damos (Ed.), *Multi-task performance* (pp. 207–216). Taylor and Francis.

Estes, S. (2015). The workload curve: Subjective mental workload. *Human Factors*, *57*(7), 1174–1187. https://doi.org/10.1177/0018720815592752

Fink, R., Hill, D., & O'Hara, J. M. (2004). *Human factors guidance for control room and digital human-system interface design and modification: Guidelines for planning, specification,*

design, licensing, implementation, training, operation and maintenance (Technical Report No. 1008122). U.S. Department of Energy. https://doi.org/10.2172/835085

Fleger, S. A. (2012, July 22). *A philosohical perspective and summary of IEEE's human factors standards on computerized operating procedure systems (COPS)*. the 8th Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies (NPIC&HMIT), San Diego, CA, United States.

Funke, M. E., Warm, J. S., Matthews, G., Riley, M., Finomore, V., Funke, G. J., Knott, B., & Vidulich, M. A. (2010). A comparison of cerebral hemovelocity and blood oxygen saturation levels during vigilance performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(18), 1345–1349. https://doi.org/10.1177/154193121005401809

Gawron, V. J. (2019). *Human performance, workload, and situational awareness measures handbook* (3rd edition). CRC Press. crcpress.com/Human-Performance-Workload-and-Situational-Awareness-Measures-Handbook/Gawron/p/book/9781138391574

Georgia Power. (2019). *Twentieth/Twenty-first semi-annual vogtle construction monitoring report* (Docket 29849). Georgia Power Company. https://www.georgiapower.com/content/dam/georgia-power/pdfs/company-pdfs/VCM-20-21-Full-Report.pdf

Gertman, D. I., Hallbert, B. P., Parrish, M. W., Sattision, M. B., Brownson, D., & Tortorelli, J. P. (2002). *Review of findings for human performance contribution to risk in operating events* (NUREG/CR-6753). U.S. Nuclear Regulatory Commission. https://www.nrc.gov/reading-rm/doc-collections/nuregs/contract/cr6753/

Gevins, A. S., Zeitlin, G. M., Doyle, J. C., Schaffer, R. E., & Callaway, E. (1979). EEG patterns during 'cognitive' tasks. II. Analysis of controlled tasks. *Electroencephalography and Clinical Neurophysiology*, *47*(6), 704–710. https://doi.org/10.1016/0013-4694(79)90297-9

Gopher, D., & Donchin, E. (1986). Workload: An examination of the concept. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance, Vol. 2: Cognitive processes and performance* (pp. 1–49). John Wiley & Sons.

GSE Power Systems. (2011). *Generic PWR nuclear plant operating manual: Loss of alternating current power to 1A-SA and 1B-SB buses*. GSE Power Systems.

Gundel, A., & Wilson, G. F. (1992). Topographical changes in the ongoing EEG related to the difficulty of mental tasks. *Brain Topography*, *5*(1), 17–25. https://doi.org/10.1007/BF01129966

Guznov, S., Reinerman-Jones, L. E., & Marble, J. (2012). Applicability of situation awareness and workload metrics for use in assessing nuclear power plant designs. In K. M. Stanney & K. S. Hale (Eds.), *Advances in cognitive engineering and neuroergonomics* (pp. 91–98). CRC Press.

Ha, J. S., Seong, P. H., Lee, M. S., & Hong, J. H. (2007). Development of human performance measures for human factors validation in the advanced MCR of APR-1400. *IEEE Transactions on Nuclear Science*, *54*(6), 2687–2700. https://doi.org/10.1109/TNS.2007.907549

Hallbert, B. P., Joe, J. C., Blackwood, L. G., Dudenhoeffer, D. D., & Hansen, K. F. (2006). *Developing human performance measures* (INL/CON-06-01256). Idaho National Laboratory. https://inldigitallibrary.inl.gov/sites/sti/sti/3394955.pdf

Hancock, P. A., & Meshkati, N. (Eds.). (1988). *Human mental workload*. North-Holland.

Hancock, P. A., & Warm, J. S. (1989). A dynamic model of stress and sustained attention. *Human Factors*, *31*(5), 519–537. https://doi.org/10.1177/001872088903100503

Hancock, P. A., Williams, G., Manning, C. M., & Miyake, S. (1995). Influence of task demand characteristics on workload and performance. *The International Journal of Aviation Psychology*, *5*(1), 63–86. https://doi.org/10.1207/s15327108ijap0501_5

Hankins, T. C., & Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation, Space, and Environmental Medicine*, *69*(4), 360–367.

Harris, J., Reinerman-Jones, L. E., & Teo, G. (2017). The impact of simulation display on nuclear power plant task error frequencies. In S. M. Cetiner, P. Fechtelkotter, & M. Legatt (Eds.), *Advances in human factors in energy: Oil, gas, nuclear and electric power industries* (pp. 133–144). Springer. https://doi.org/10.1007/978-3-319-41950-3_12

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in psychology* (Vol. 52, pp. 139–183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Helton, W. S., Warm, J. S., Tripp, L. D., Matthews, G., Parasuraman, R., & Hancock, P. A. (2010). Cerebral lateralization of vigilance: A function of task difficulty. *Neuropsychologia*, *48*(6), 1683–1688. https://doi.org/10.1016/j.neuropsychologia.2010.02.014

Henelius, A., Hirvonen, K., Holm, A., Korpela, J., & Muller, K. (2009). Mental workload classification using heart rate metrics. *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1836–1839. https://doi.org/10.1109/IEMBS.2009.5332602

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology*, *45*(1–3), 73–93. https://doi.org/10.1016/S0301-0511(96)05223-4

Hughes, N., & D'Agostino, A. (2016). Gathering meaningful data from novices and or in simplified operating environments to inform us about highly complex operational environments. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 47–50. https://doi.org/10.1177/1541931213601011

Hughes, N., D'Agostino, A., & Reinerman-Jones, L. E. (2017). The NRC Human Performance Test Facility: An approach to data collection using novices and a simplified environment. In S. M. Cetiner, P. Fechtelkotter, & M. Legatt (Eds.), *Advances in human factors in energy: Oil, gas, nuclear and electric power industries* (pp. 183–192). Springer. https://doi.org/10.1007/978-3-319-41950-3_16

Hugo, J. V., Slay III, L., & Hernandez, J. (2017, June 11). *Human factors and modeling methods in the development of control room modernization concepts*. the 10th Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies (NPIC&HMIT), San Francisco, CA, United States. https://www.osti.gov/servlets/purl/1364031

Hulbert, T. (1989). *A comparison of the "NASA-TLX" and "ISA" subjective workload rating techniques* [Internal Report]. Civil Aviation Authority Air Traffic Control Evaluation Unit.

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

Ikuma, L. H., Harvey, C., Taylor, C. F., & Handal, C. (2014). A guide for assessing control room operator performance using speed and accuracy, perceived workload, situation awareness, and eye tracking. *Journal of Loss Prevention in the Process Industries*, *32*, 454–465. https://doi.org/10.1016/j.jlp.2014.11.001

International Organization for Standardization. (2004). *Ergonomic principles related to mental workload — Part 3: Principles and requirements concerning methods for measuring and assessing mental workload* (ISO 10075-3:2004). International Organization for Standardization. https://www.iso.org/standard/27571.html

International Organization for Standardization. (2013). *Ergonomic design of control centres — Part 4: Layout and dimensions of workstations* (ISO 11064-4:2013). International Organization for Standardization. https://www.iso.org/standard/54419.html

Joe, J. C., & Boring, R. L. (2017). Using the human systems simulation laboratory at Idaho National Laboratory for safety focused research. In S. M. Cetiner, P. Fechtelkotter, & M. Legatt (Eds.), *Advances in human fhactors in energy: Oil, gas, nuclear and electric power industries* (pp. 193–201). Springer. https://doi.org/10.1007/978-3-319-41950-3_17

Joe, J. C., Boring, R. L., & Persensky, J. J. (2012). *Commercial utility perspectives on nuclear power plant control room modernization*. 2039–2046. https://www.osti.gov/biblio/1056001

Jordan, C. S. (1992). *Experimental study of the effects of an instantaneous self-assessment workload recorder on task performance* (DRA Technical Memorandum CAD5 92011). Defence Research Agency Maritime Command Control Division.

Jorna, P. G. A. M. (1993). Heart rate and workload variations in actual and simulated flight. *Ergonomics*, *36*(9), 1043–1054. https://doi.org/10.1080/00140139308967976

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.

Kamzanova, A. T., Kustubayeva, A. M., & Matthews, G. (2014). Use of EEG workload indices for diagnostic monitoring of vigilance decrement. *Human Factors*, *56*(6), 1136–1149. https://doi.org/10.1177/0018720814526617

Kim, S., Park, J., Han, S., & Kim, H. (2010). Development of extended speech act coding scheme to observe communication characteristics of human operators of nuclear power plants under abnormal conditions. *Journal of Loss Prevention in the Process Industries*, *23*(4), 539–548. https://doi.org/10.1016/j.jlp.2010.04.005

Kirwan, B., & Ainsworth, L.K. (Eds.). (1992). A Guide to Task Analysis: The Task Analysis Working Group (1st ed.). CRC Press. https://doi.org/10.1201/b16826

Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. In D. L. Damos (Ed.), *Multi-task performance* (pp. 279–328). Taylor and Francis.

Kramer, A. F., Sirevaag, E. J., & Braune, R. (1987). A psychophysiological assessment of operator workload during simulated flight missions. *Human Factors*, *29*(2), 145–160. https://doi.org/10.1177/001872088702900203

Kubota, T., Fang, J., Guan, Z., Brown, R. A., & Krueger, J. M. (2001). Vagotomy attenuates tumor necrosis factor-α-induced sleep and EEG δ-activity in rats. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *280*(4), R1213–R1220. https://doi.org/10.1152/ajpregu.2001.280.4.R1213

Kurimori, S., & Kakizaki, T. (1995). Evaluation of work stress using psychological and physiological measures of mental activity in a paced calculating task. *Industrial Health*, *33*(1), 7–22. https://doi.org/10.2486/indhealth.33.7

Lackey, S. J., Reinerman-Jones, L. E., & Salcedo, J. N. (2014, April 15). *Equal but different: 5 research strategies for improving conclusions drawn from novice populations*. the 2014 MODSIM World Conference, Hampton, VA, United States. https://pdfs.semanticscholar.org/c048/57b270b234aa0ada4737bf0af2535dc0622c.pdf

Leis, R., & Reinerman-Jones, L. E. (2015). Methodological implications of confederate use for experimentation in safety-critical domains. *Procedia Manufacturing*, *3*, 1233–1240. https://doi.org/10.1016/j.promfg.2015.07.258

Leis, R., Reinerman-Jones, L. E., Mercado, J., Barber, D., & Sollins, B. (2014). Workload from nuclear power plant task types across repeated sessions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 210–214. https://doi.org/10.1177/1541931214581044

Lew, R., Ulrich, T. A., & Boring, R. L. (2017). Nuclear reactor crew evaluation of a computerized operator support system HMI for chemical and volume control system. In D. D. Schmorrow

& C. M. Fidopiastis (Eds.), *Augmented cognition. Enhancing cognition and behavior in complex human environments* (pp. 501–513). Springer International Publishing. https://doi.org/10.1007/978-3-319-58625-0_36

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

Lin, C. J., Yenn, T.-C., & Yang, C.-W. (2010). Automation design in advanced control rooms of the modernized nuclear power plants. *Safety Science*, *48*(1), 63–71. https://doi.org/10.1016/j.ssci.2009.05.005

Manusov, V. (Ed.). (2005). *The sourcebook of nonverbal measures: Going beyond words*. Routledge.

Matthews, G., & Reinerman-Jones, L. E. (2017). *Workload assessment: How to diagnose workload issues and enhance performance*. Human Factors and Ergonomics Society.

Matthews, G., Reinerman-Jones, L. E., Barber, D., Teo, G., Wohleber, R. W., Lin, J., & Panganiban, A. R. (2016). Resilient autonomous systems: Challenges and solutions. *2016 Resilience Week (RWS)*, 208–213. https://doi.org/10.1109/RWEEK.2016.7573335

Matthews, G., Reinerman-Jones, L. E., Wohleber, R. W., Lin, J., Mercado, J., & Abich IV., J. (2015). Workload is multidimensional, not unitary: What now? In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Foundations of Augmented Cognition* (pp. 44–55). Springer.

Matthews, G., Wohleber, R. W., & Lin, J. (2019). Stress, skilled performance, and expertise: Overload and beyond. In P. Ward, J. M. Schraagen, J. Gore, & E. M. Roth (Eds.), *The Oxford Handbook of Expertise* (pp. 490–524). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198795872.013.22

Megaw, T. (2005). The definition and measurement of mental workload. In J. R. Wilson & N. Corlett (Eds.), *Evaluation of human work* (3rd edition, pp. 521–551). CRC Press.

Mercado, J. (2014). *Assessing the effectiveness of workload measures in the nuclear domain* [Doctoral dissertation, University of Central Florida]. https://stars.library.ucf.edu/etd/1287

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Min, D., Chung, Y. H., & Yoon, W. C. (2004, September). *Comparative analysis of communication at main control rooms of nuclear power plants*. The IFAC/ IFIP/IFORS/IEA symposium, Atlanta, GA, United States.

Miyake, S. (2001). Multivariate workload evaluation combining physiological and subjective measures. *International Journal of Psychophysiology*, *40*(3), 233–238. https://doi.org/10.1016/S0167-8760(00)00191-4

Moray, N. (1967). Where is capacity limited? A survey and a model. *Acta Psychologica*, *27*, 84–92. https://doi.org/10.1016/0001-6918(67)90048-0

Moray, N. (1979). Models and measures of mental workload. In N. Moray (Ed.), *Mental workload: Its theory and measurement* (pp. 13–22). Plenum Press.

Moray, N. (Ed.). (2013). *Mental workload: Its theory and measurement*. Springer. https://doi.org/10.1007/978-1-4757-0884-4

Mosleh, A., & Chang, Y. H. (2004). Model-based human reliability analysis: prospects and requirements. *Reliability Engineering & System Safety*, *83*(2), 241–253. https://doi.org/10.1016/j.ress.2003.09.014

Mulder, L. (Ben) J. M., de Waard, D., & Brookhuis, K. A. (2004). Estimating mental effort using heart rate and heart rate variability. In N. Stanton, A. Hedge, K. A. Brookhuis, E. Salas, & H. Hendrick (Eds.), *Handbook of Human Factors and Ergonomics Methods* (pp. 201–208). CRC Press.

O'Hara, J. M., Brown, W. S., Lewis, P. M., & Persensky, J. J. (2002). *Human-system interface design review guidelines* (NUREG-0700, Revision 2). U.S. Nuclear Regulatory Commission. https://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr0700/

O'Hara, J. M., & Higgins, J. C. (2010). *Human-system interfaces to automatic systems: Review guidance and technical basis* (BNL-91017-2010). Brookhaven National Laboratory. https://www.bnl.gov/isd/documents/71082.pdf

O'Hara, J. M., Higgins, J. C., Brown, W. S., Fink, R., Persensky, J. J., Lewis, P. M., Kramer, J., & Szabo, A. (2008). *Human factors considerations with respect to emerging technology in nuclear power plants* (NUREG/CR-6947). U.S. Nuclear Regulatory Commission. https://www.nrc.gov/reading-rm/doc-collections/nuregs/contract/cr6947/

O'Hara, J. M., Higgins, J. C., Fleger, S. A., & Barnes, V. (2010). Guidance for human-system interfaces to automatic systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(4), 403–407. https://doi.org/10.1177/154193121005400428

O'Hara, J. M., Higgins, J. C., Fleger, S. A., & Pieringer, P. A. (2012). *Human factors engineering program review model* (NUREG-0711, Revision 3). U.S. Nuclear Regulatory Commission. https://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr0711/

Patel, S. (October, 2022). China Will Add Two More AP1000 Nuclear Reactors. *Power*, https://www.powermag.com/china-will-add-two-more-ap1000-nuclear-reactors/

Prinzel, L. J., Freeman, F. G., Scerbo, M. W., Mikulka, P. J., & Pope, A. T. (2009). A closed-loop system for examining psychophysiological measures for adaptive task allocation. *The International Journal of Aviation Psychology*. https://doi.org/10.1207/S15327108IJAP1004_6

Raeisi, S., Osqueizadeh, R., Maghsoudipour, M., & Jafarpisheh, A. S. (2016). Ergonomic redesign of an industrial control panel. *The International Journal of Occupational and Environmental Medicine*, *7*(3), 186–192. https://doi.org/10.15171/ijoem.2016.756

Ragan, E. D., Bowman, D. A., Kopper, R., Stinson, C., Scerbo, S., & McMahan, R. P. (2015). Effects of field of view and visual complexity on virtual reality training effectiveness for a visual scanning task. *IEEE Transactions on Visualization and Computer Graphics*, *21*(7), 794–807. https://doi.org/10.1109/TVCG.2015.2403312

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-13*(3), 257–266. https://doi.org/10.1109/TSMC.1983.6313160

Reason, J. (1995). Understanding adverse events: human factors. *BMJ Quality & Safety*, *4*(2), 80–89. https://doi.org/10.1136/qshc.4.2.80

Reinerman-Jones, L. E., Cosenzo, K., & Nicholson, D. (2010). Subjective and objective measures of operator state in automated systems. In T. Marek, W. Karwowski, & V. Rice (Eds.), *Advances in understanding human performance: Neuroergonomics, human factors design, and special populations* (pp. 122–131). CRC Press.

Reinerman-Jones, L. E., Guznov, S., Mercado, J., & D'Agostino, A. (2013). Developing methodology for experimentation using a nuclear power plant simulator. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Foundations of augmented cognition* (pp. 181–188). Springer. https://doi.org/10.1007/978-3-642-39454-6_19

Reinerman-Jones, L. E., Guznov, S., Tyson, J., D'Agostino, A., & Hughes, N. (2015). *Workload, situation awareness, and teamwork* (NUREG/CR-7190). U.S. Nuclear Regulatory Commission. https://www.nrc.gov/reading-rm/doc-collections/nuregs/contract/cr7190/

Reinerman-Jones, L. E., Harris, J., Hughes, N., & D'Agostino, A. (2017, June 11). *Workload response to soft controls presented on two interfaces*. the 10th Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies (NPIC&HMIT), San Francisco, CA, United States. http://npic-hmit2017.org/wp-content/data/pdfs/382-21353.pdf

Reinerman-Jones, L. E., Hughes, N., D'Agostino, A., & Matthews, G. (2019). Human performance metrics for the nuclear domain: A tool for evaluating measures of workload, situation awareness and teamwork. *International Journal of Industrial Ergonomics*, *69*, 217–227. https://doi.org/10.1016/j.ergon.2018.12.001

Reinerman-Jones, L. E., Lin, J., Barber, D., Matthews, G., & Hughes, N. (2019). Multi-dimensional workload in two types of nuclear power plant simulators. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 1777–1781. https://doi.org/10.1177/1071181319631432

Reinerman-Jones, L. E., Matthews, G., Barber, D. J., & Abich IV., J. (2014). Psychophysiological metrics for workload are demand-sensitive but multifactorial. *Proceedings of the Human*

Factors and Ergonomics Society Annual Meeting, *58*(1), 974–978. https://doi.org/10.1177/1541931214581204

Reinerman-Jones, L. E., Matthews, G., Langheim, L. K., & Warm, J. S. (2011). Selection for vigilance assignments: a review and proposed new direction. *Theoretical Issues in Ergonomics Science*, *12*(4), 273–296. https://doi.org/10.1080/14639221003622620

Reinerman-Jones, L. E., Matthews, G., Warm, J. S., Langheim, L. K., Parsons, K., Proctor, C. A., Siraj, T., Tripp, L. D., & Stutz, R. M. (2006). Cerebral blood flow velocity and task engagement as predictors of vigilance performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(12), 1254–1258. https://doi.org/10.1177/154193120605001210

Reinerman-Jones, L. E., & Mercado, J. (2014). *Human performance test facility task order 1 technical report* (JCN # V621). U.S. Nuclear Regulatory Commission.

Roscoe, A. H. (1992). Assessing pilot workload. Why measure heart rate, HRV and respiration? *Biological Psychology*, *34*(2), 259–287. https://doi.org/10.1016/0301-0511(92)90018-P

Roscoe, A. H. (1993). Heart rate as a psychophysiological measure for in-flight workload assessment. *Ergonomics*, *36*(9), 1055–1062. https://doi.org/10.1080/00140139308967977

Savchenko, K., Medema, H., Boring, R. L., & Ulrich, T. (2018). Comparison of mutual awareness in analog vs. digital control rooms. In R. L. Boring (Ed.), *Advances in Human Error, Reliability, Resilience, and Performance* (pp. 192–199). Springer. https://doi.org/10.1007/978-3-319-60645-3_19

Sears, A., & Shneiderman, B. (1991). High precision touchscreens: design strategies and comparisons with a mouse. *International Journal of Man-Machine Studies*, *34*(4), 593–613. https://doi.org/10.1016/0020-7373(91)90037-8

Slater-Thompson, N. (2014). *Nuclear Regulatory Commission resumes license renewals for nuclear power plants*. U.S. Energy Information Administration. https://www.eia.gov/todayinenergy/detail.php?id=18591

Spielman, Z., & Hill, R. (2017, June 11). *A summary comparison of design evaluation techniques*. the 10th International Topical Meeting on Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies (NPIC&HMIT), San Francisco, CA, United States. https://www.osti.gov/servlets/purl/1378447

Szalma, J. L. (2014). On the application of motivation theory to human factors/ergonomics: Motivational design principles for human–technology interaction. *Human Factors*, *56*(8), 1453–1471. https://doi.org/10.1177/0018720814553471

Tanner Jr., W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, *61*(6), 401–409. https://doi.org/10.1037/h0058700

Tattersall, A. J., & Foord, P. S. (1996). An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, *39*(5), 740–748. https://doi.org/10.1080/00140139608964495

Taylor, G. (2012). *Comparing types of adaptive automation within a multi-tasking environment* [Doctoral dissertation]. https://stars.library.ucf.edu/etd/2321/

Taylor, G., Reinerman-Jones, L. E., Cosenzo, K., & Nicholson, D. (2010). Comparison of multiple physiological sensors to classify operator state in adaptive automation systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(3), 195–199. https://doi.org/10.1177/154193121005400302

Tran, T. Q., Boring, R. L., Dudenhoeffer, D. D., Hallbert, B. P., Keller, M. D., & Anderson, T. M. (2007). Advantages and disadvantages of physiological assessment for next generation control room design. *2007 IEEE 8th Human Factors and Power Plants and HPRCT 13th Annual Meeting*, 259–263. https://doi.org/10.1109/HFPP.2007.4413216

Turner, J. R. (1994). *Cardiovascular reactivity and stress: Patterns of physiological response*. Plenum Press.

Ulrich, T. A., Boring, R. L., & Lew, R. (2015, August). Control board digital interface input devices-touchscreen, trackpad, or mouse?. *In 2015 Resilience Week (RWS)* (pp. 1-6). IEEE.

U.S. Department of Energy. (2009). *Human performance improvement handbook: Human performance tools for individuals, work teams and management* (DOE-HDBK-1028-2009). U.S. Department of Energy. https://www.standards.doe.gov/files/doe-hdbk-1028-2009-human-performance-improvement-handbook-volume-2-human-performance-tools-for-individuals-work-teams-and-management

U.S. Nuclear Regulatory Commission. (2006). Staff Requirements – Meeting with Advisory Committee on Reactor Safeguards (SRM-M061020). U.S. Nuclear Regulator Commission. https://adamsxt.nrc.gov/navigator/AdamsXT/content/downloadContent.faces?objectStoreName=MainLibrary&ForceBrowserDownloadMgrPrompt=false&vsId=%7b440AF45F-D515-452A-BD02-245777EF9548%7d

U.S. Nuclear Regulatory Commission. (2008). *Guidance to operators at the controls and to senior operators in the control room of a nuclear power unit* (REGULATORY GUIDE 1.114). U.S. Nuclear Regulatory Commission. https://www.nrc.gov/docs/ML0823/ML082380236.pdf

U.S. Nuclear Regulatory Commission. (2008). Analysis of options and recommendations for new reactor simulator training of NRC inspectors (SECY-08-0195). U.S. Nuclear Regulatory Commission. https://adamsxt.nrc.gov/navigator/AdamsXT/packagecontent/packageContent.faces?id={BE294C07-D4F8-4030-8884-E67C9939CD85}&objectStoreName=MainLibrary&wId=1669916626294

U.S. Nuclear Regulatory Commission. (2009). Staff Requirements – Briefing on risk-informed, performance-based regulations (SRM-M090204B). U.S. Nuclear Regulatory Commission.

https://adamsxt.nrc.gov/navigator/AdamsXT/content/downloadContent.faces?objectStoreName=MainLibrary&ForceBrowserDownloadMgrPrompt=false&vsId=%7bA48838E6-D0A2-4969-B4CE-F168068567B6%7d

U.S. Nuclear Regulatory Commission. (2016). *Standard Review Plan for the Review of Safety Analysis Reports for Nuclear Power Plants: LWR Edition - Human Factors Engineering (NUREG-0800, Chapter 18)* (NUREG-0800, Chapter 18). U.S. Nuclear Regulatory Commission. https://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr0800/ch18/index.html

U.S. Nuclear Regulatory Commission. (2017). *NRC Regulations Title 10, Code of Federal Regulations*. U.S. Nuclear Regulatory Commission. https://www.nrc.gov/reading-rm/doc-collections/cfr/

U.S. Nuclear Regulatory Commission. (2021). *2021-2022 information digest* (NUREG-1350). https://www.nrc.gov/docs/ML1924/ML19242D326.pdf

Veltman, J. A., & Gaillard, A. W. K. (1996). Physiological indices of workload in a simulated flight task. *Biological Psychology*, *42*(3), 323–342.

Warm, J. S., Dember, W. N., & Hancock, P. A. (1996). Vigilance and workload in automated systems. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 183–200). Lawrence Erlbaum Associates.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human factors*, *50*(3), 433-441.

Warm, J. S., Tripp, L. D., Matthews, G., & Helton, W. S. (2012). Cerebral hemodynamic indices of operator fatigue in vigilance. In G. Matthews, P. A. Desmond, C. Neubauer, & P. A. Hancock (Eds.), *The Handbook of Operator Fatigue* (pp. 197–207). CRC Press. https://doi.org/10.1201/9781315557366-13

Webster Jr., M., & Sell, J. (Eds.). (2007). *Laboratory Experiments in the Social Sciences*. Academic Press.

Wertheim, A. H. (1981). Occipital alpha activity as a measure of retinal involvement in oculomotor control. *Psychophysiology*, *18*(4), 432–439. https://doi.org/10.1111/j.1469-8986.1981.tb02476.x

Whaley, A. M., Xing, J., Boring, R. L., Hendrickson, S. M. L., Joe, J. C., & Le Blanc, K. L. (2012). *Building a psychological foundation for human reliability analysis,* (NUREG-2114). U.S. Nuclear Regulatory Commission. https://www.nrc.gov/docs/ML1131/ML113180490.pdf

Whaley, A. M., Xing, J., Boring, R. L., Hendrickson, S. M. L., Joe, J. C., Le Blanc, K. L., & Morrow, S. L. (2016). *Cognitive basis for human reliability analysis* (NUREG-2114). U.S. Nuclear Regulatory Commission. https://www.nrc.gov/docs/ML1601/ML16014A045.pdf

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, *50*(3), 449–455. https://doi.org/10.1518/001872008X288394

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2015). *Engineering psychology and human performance* (4th edition). Routledge.

Williges, R. C., & Wierwille, W. W. (1979). Behavioral measures of aircrew mental workload. *Human Factors*, *21*(5), 549–574. https://doi.org/10.1177/001872087902100503

Wilson, G. F. (2002). An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology*, *12*(1), 3–18. https://doi.org/10.1207/S15327108IJAP1201_2

Wilson, G. F., Fullenkamp, P., & Davis, I. (1994). Evoked-potential, cardiac, blink, and respiration measures of pilot workload in air-to-ground missions. *Aviation, Space, and Environmental Medicine*, *65*(2), 100–105.

Xing, J., Chang, Y. J., & DeJesus, J. (2020). *NUREG-2198 The General Methodology of an Integrated Human Event Analysis System (IDHEAS-G)* (NUREG-2198). U.S. Nuclear Regulatory Commission. ML20329A428

Xu, J., Anders, S., Pruttianan, A., France, D. J., Lau, N., Adams, J. A., & Weinger, M. B. (2017, June 11). *Human performance measures for the evaluation of nuclear power plant control room interfaces: A systematic review*. the 10th International Topical Meeting on Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies (NPIC&HMIT), San Francisco, CA, United States. http://npic-hmit2017.org/wp-content/data/pdfs/265-19916.pdf

Yeh, Y.-Y., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors*, *30*(1), 111–120. https://doi.org/10.1177/001872088803000110

APPENDIX A
SIMULATED ENVIRONMENTS

Work has been done in the NPP domain to understand the types of tasks operators perform, but systematically investigating and measuring operator performance, errors, and states in a highly controlled experimental setting while executing those tasks has been limited. Developing an appropriate experimental methodology is necessary to effectively evaluate questions concerned with the factors that influence operators' performance, errors, and states. The first step for developing an appropriate methodology is understanding the environment with which experiments could occur. The place for conducting experiments that comes to mind most quickly is in an actual MCR at an NPP. However, this is not a realistic, safe, or feasible option for experimental research. Therefore, simulated environments are the best alternative to an online operational plant. Simulated environments offer a host of advantages in comparison to real-world testing, such as reduced costs associated with developing, running, and maintaining these systems, consistency, and control of variable manipulations, logging capabilities for real-time and post-hoc analysis, and increased safety for consequences resulting from operator or system error. This list of benefits leads one to believe that all simulated environments are created equal. However, it is important to consider the various types of simulators available.

A.1 Full-Scale Simulated Environments

The majority of NPP simulators have been developed for training purposes and are located in training centers. These simulators are full-scale copies of MCRs including full size instrument panels of a particular NPP (Figure A-1 Example of a full-scale MCR training simulator for an NPP). The physics of the processes is also preserved, meaning the simulator responds as if a physical plant were feeding the controls and thus the dynamically yoked systems affect one another accordingly. Several challenges are associated with using training simulators for research purposes. One challenge is that the scenarios and tasks used in training are complex and intended for use by licensed operators. However, access to operators is limited for use in experimentation. Therefore, scenarios and tasks used in experimentation need consideration in terms of the experience of the sample. Another issue with training simulators is the capability to record important human performance indices such as response time, response accuracy, and team communications. Further, the training simulators are often extremely large and require separate housing and maintenance crews. Finally, reconfiguring the simulator is not feasible since the simulators are usually created to represent a specific NPP.

Figure A-1 Example of a full-scale MCR training simulator for an NPP

A.2 Mixed Reality Environments

An alternative to full-scale simulators is mixed reality simulators. These environments include parts of a live environment and parts of a virtual environment. In other words, controls might be real so that the person feels as though they are in the actual physical environment, but the tasks are virtual three-dimensional (3D) figures on a computer display. An example of this is illustrated in Figure A-2 Example of a mixed-reality environment (right) simulating astronauts experience in. The picture on the left is an astronaut in a shuttle in space, whereas the picture on the left is a participant in a mixed-reality space capsule viewing earth through virtual portals. You can see that the physical environment is partially replicated in terms of close quarters, a computer screen in front, a clipboard for paperwork, and small windows. Not seen, is the sound effects associated with takeoff and landing. The virtual environment is presented through the small portals and uses a real picture of the earth with graphic effects overlaid to enhance the image in order to most closely resemble that seen by the astronauts. Mixed reality environments offer a less expensive format to train people and test experimental hypotheses. There is more control over configuring this type of a simulator and there is a high level of immersion or a feeling of being absorbed. Also, logging capabilities for the timing of events and participant responses are easier to build and integrate. However, in an effort to replicate the physical environment and task experience as closely as possible to the real thing, researchers lose control over every factor that could be influencing the person's performance, errors, or states. In the space example, a person's responses could be a result of the shuttle construct, the sounds, or the virtual visual stimuli. If research on each of these components did not occur independently prior to putting them together in the mixed reality simulator, then only general relationships about a person's experience can be made. Causation statements are cautioned.

Figure A-2 Example of a mixed-reality environment (right) simulating astronauts experience in space.

A.3 Cave Automatic Virtual Environments (CAVE)

Another type of highly immersive environment is a cave automatic virtual environment (CAVE, Figure A-3 Example of a CAVE). CAVEs usually have multiple projectors that display images on large screens or walls. These environments can vary in height and width, but an example is the Northrup Grumman Virtual Immersive Portable Environment (VIPE), which is 7 ft tall with a 120-degree viewing area. The VIPE can have 180- or 360-degree viewing areas also depending on the number of screens. These environments can project virtually any image from various gaming and research engines, as well as DVDs or blue rays. CAVEs enable high fidelity virtual reality with experimental control over factors influencing a person's experience. Speakers and other sensors like the connect can be integrated if the research question requires. Maintenance for such a system is minimal, but a large space is needed for housing and appropriate ventilation is needed to keep the projectors cool. The cost to purchase these systems varies but can be upwards of $100,000.

Figure A-3 Example of a CAVE

A.4 Computer Simulated Environments

Computer simulators provide a good balance between the fidelity of the task environment and cost. However, in order to preserve the ecological validity of the results obtained from such simulators, several characteristics are important to consider. One of them is the fidelity of the simulator; the simulator needs to have sufficient fidelity to give the researcher sufficient confidence that the results observed in the simulator would be observed in the real environment. For an NPP simulator, this means that the instrument panels and their components (e.g., knobs, switches, and dials) need to be similar to the ones in actual panels. Another important component of a suitable simulator is its ability to accommodate a variety of scenarios within the domain under investigation. An NPP simulator should be capable of control scenarios for both normal states of the plant such as start-up and shut-down and off-normal events (e.g., LOCA). In addition, a task that requires teams or crew needs to use a simulator capable of simulating the teamwork environment with built-in interdependency of team member actions. This is the case for NPP operation and therefore a computer-based MCR simulator should be physics-based, such that responses to events that require actions at one panel result in changes at other panels. Based upon these considerations, the GSE NPP simulator was selected as the simulation platform for developing our experimental methodology. This is a desktop simulator of a Westinghouse AP1000 (Figure A-4). The hardware set-up consists of four desktop computers and eight monitors, one computer and two monitors are designated as the instructor or researcher station and the others show fully-functional panels and can be configured for individual and team studies. The simulator also allows accommodating various changes to the panels and has extensive data recording capabilities.

Figure A-4 Westinghouse AP1000 simulator

APPENDIX B
PARTICIPANT TRAINING

B.1 Participant Training Phase 1: Three-way Communication Skills

Participant Training Phase 1: 3-way communication skills

Please read carefully and communicate instructions using the 3-way communication method.

Remember - if you are unsure about an element or believe that there is a misunderstanding in the communication between you and the SRO clarify by addressing them and asking them to repeat the previous action. Your cue to communicate is bolded below.

The instructor will give you feedback at the end of each scenario. Remember to do your best

3-way communication Practice Scenario 1:

Instructions: *Respond to the instructions given to you by the SRO*

- Listen carefully as the SRO instructs, you to verify that valve PCV-445B is shut.
- **RO1 (You), acknowledge directions and respond to the SRO.**
- Listen as the SRO confirms your response.
- **RO1 (You), reply to SRO verify the state of PCV-445B.**
- Listen as the SRO acknowledges and responds to you.
- **RO1 (You), provide confirmation to SRO.**

3-way communication Practice Scenario 2:

Instructions: *Respond as if you did not hear the command and need further clarification*

- Listen as SRO instructs you to shut valve 1CS-8.
- **RO1 (You)**, **respond *as if you did not hear the command and need further clarification***.
- Listen as SRO repeats directive.
- **RO1 (You), acknowledge directions and respond to the SRO.**
- Listen as the SRO confirms your response.
- **RO1 (You), reply to SRO verify that 1CS-8 is shut.**
- Listen as the SRO acknowledges and responds to you.
- **RO1, provide confirmation to SRO.**

3-way communication Practice Scenario 3:

Instructions: *Respond as if you cannot find the location of the control.*

- Listen as SRO instructs you to verify and report back the state of gauge PI-403.1.
- **RO1 (You), acknowledge directions and respond to the SRO.**
- Listen as the SRO confirms your response.
- **RO1 (You), Pause for a moment to simulate looking for the control. Then respond *as if you cannot find the control and need further clarification***.

B.2 Participant Training Phase 2: Instruments and Controls Type Evaluation

Participant Training Phase 2: Instruments and Control Types Evaluation

Date_____

Participant_____

Instructor_____

The following tool is designed to assist in the evaluation of participants after completing Phase2. Include any notes in the Observation Score column.

Scoring rubric:
0 = Not observed
1 = Partial completion, with significant omissions and/or errors
2 = Generally complete, with minor omissions and/or errors

Instruments and Control Types Evaluation Scenario 1:

Read: You will be evaluated on three things:

1) Locating the item.
2) Correctly assessing the state of the item.
3) Following directions to change the state of the plant by verifying gauge level changes, checking lights, or flipping a switch.

Do your best to use the 3-way communication we practiced in the previous phase, but focus on locating and reporting the state, as those are the only things you are being evaluated on right now. Follow along with the information provided.

Instruction Given	Example Response	Participant Score	Feedback Given
Step 1: Checking (Valve status) **SRO, instruct RO1 to check if valve PCV-445A is shut**	SRO: R-O-1, verify valve P-C-V-4-4-5-A is shut RO1: S-R-O, understood verifying valve P-C-V-4-4-5-A is shut SRO: R-O-1, that is correct RO1: S-R-O, valve P-C-V-4-4-5-A is shut SRO: R-O-1, understood valve P-C-V-4-4-5-A is shut RO1: S-R-O, that is correct	Δ Correctly identified the location of the item of focus Δ Correctly identified the state of the item of focus Score:_____	

B.3 Participant Training Phase 3: Locating Controls Evaluation

Date_____

Participant Training Phase 3: Locating Controls Evaluation

Participant

Instructor_____

The following tool is designed to assist in the evaluation of participants after completing Phase3. Include any notes in the Observation Score column.

Please Read: The objective of this evaluation is to determine the skill level of your 3-way communication and to effectively locate specific controls within the simulator. In the first block, you may use the red arrows to guide you if you need help. Instructor feedback will be provided.

Scoring rubric:

0 = Not observed
1 = Partial completion, with significant omissions and/or errors
2 = Generally complete, with minor omissions and/or errors

Scenario Completion Evaluation Block 1:

SRO Tips	Instructions	Participant Score	Feedback Given
Who: Reactor Operator 1 **Panel: C1Mod** **Additional Info: The light box is located in the middle** **Reference the arrow**	SRO: R-O-1, verify light box R-T-A is open RO1: S-R-O, understood verifying light box R-T-A is open SRO: R-O-1, that is correct RO1: S-R-O, light box R-T-A is open SRO: R-O-1, understood light box R-T-A is open RO1: S-R-O, that is correct	Δ Gave effective 3-way communication Δ Correctly identified the location of the item of focus Δ Correctly identified the state of the item of focus Score:_____	
Who: Reactor Operator 1 **Panel: A2Mod** **Additional Info: The valve is located in the bottom left of the panel** **Reference the arrow**	SRO: R-O-1, shut valve 1-C-S-7 RO1: S-R-O, understood shutting valve 1-C-S-7 SRO: R-O-1, that is correct (RO1 Shut without speaking until shut) RO1: S-R-O, valve 1-C-S-7 is shut SRO: R-O-1, understood valve 1-C-S-7 is shut	Δ Gave effective 3-way communication Δ Correctly identified the location of the item of focus Δ Correctly identified the state of the item of focus Δ Correctly changed the state of the control Score:_____	

B.4 Participant Training Phase 4: Scenario Completion

Participant Training Phase 4: Scenario Completion

Read to participants: We will now practice all the elements together with your team member. You will also practice the ISA questionnaire. During this practice, you will hear a voice asking you to rate your workload. Please respond with a score of 1-5 as instructed. You will not be provided with any prompts or visuals. This will be your last practice before we begin the experimental task, so if you need further clarification please make sure to ask now.

Scenario Completion Practice:

(3a) Check PRZ PORV PCV -445A-SHUT: A2 Bottom Right

SRO: R-O-1, verify valve P-C-V-4-4-5-A is shut

RO1: S-R-O, understood verifying valve P-C-V-4-4-5-A is shut

SRO: R-O-1, that is correct

RO1: S-R-O, valve P-C-V-4-4-5-A is shut

SRO: R-O-1, understood valve P-C-V-4-4-5-A is shut

RO1: S-R-O, that is correct

Wait for ISA Prompt

(3b) Check letdown isolation valve 1CS-1 LCV-459- SHUT: Panel A2 Bottom Left

SRO: R-O-2 verify valve L-C-V-4-5-9 is shut

RO2: S-R-O, understood verifying valve L-C-V-4-5-9 is shut

SRO: R-O-2, that is correct

RO2: S-R-O, valve L-C-V-4-5-9 is open

SRO: R-O-2, understood valve L-C-V-4-5-9 is open

RO2: S-R-O, that is correct

Wait for ISA Prompt

(3b.1) Shut orifice isolation valve 1CS-8:A2 Bottom Left

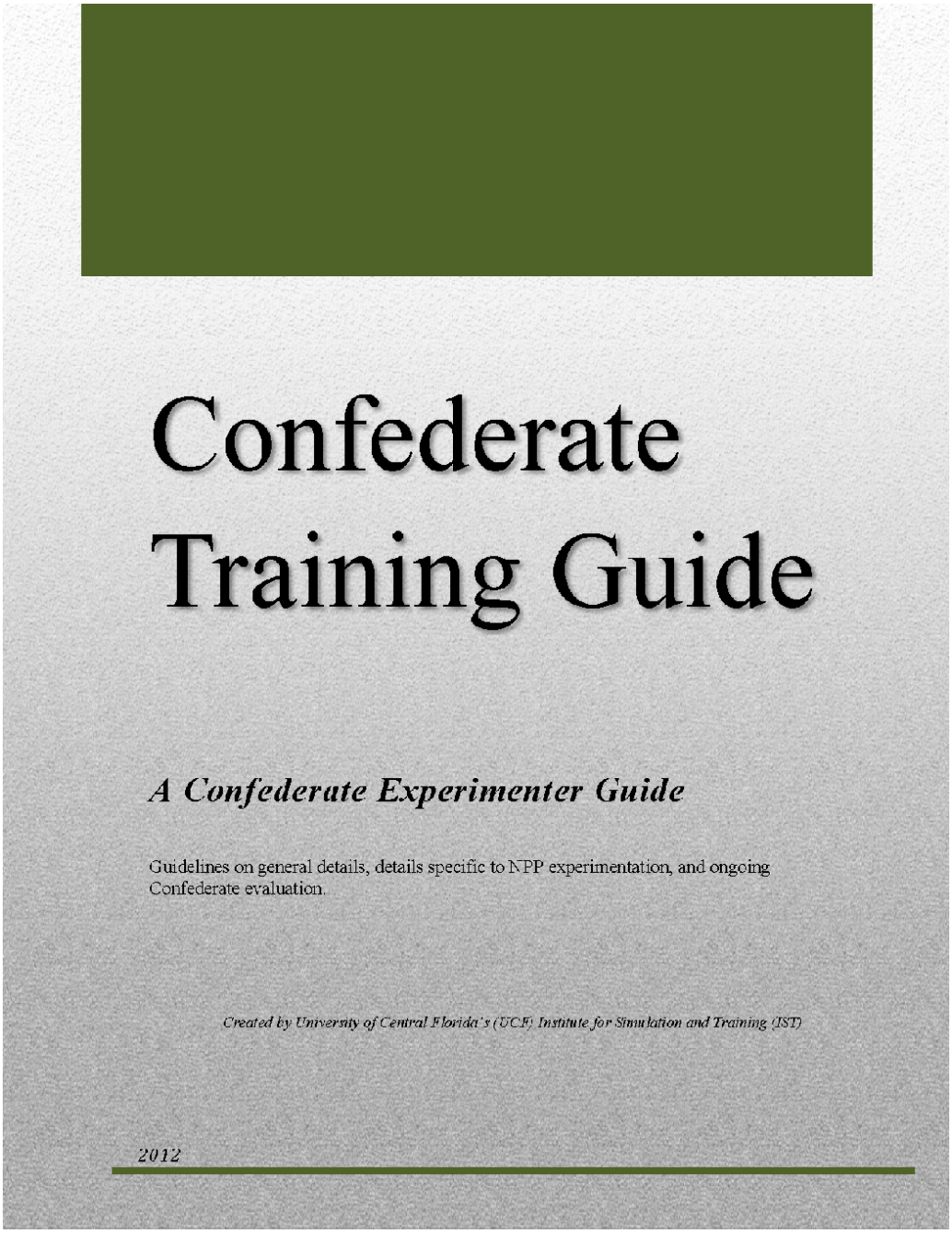
SRO: R-O-1, shut valve 1-C-S-8

APPENDIX C
SUMMARY OF PARTICIPANT TRAINING ON TWO INTERFACE GROUPS

Phase	*Desktop Interface*	*Touchscreen Interface*
Start of training	n=83	n=73
Phase 1	Score of those passed: 96.76% (n=82) Score of those failed*: 62.5% (n=1)	Score of those passed: 97.11% (n=71) Score of those failed*: 57.31% (n=2)
Phase 2	Score of those passed: 98.63 (n=82) No failures	Score of those passed: 95.68% (n=71) : 67 passed on 1st try; 4 passed on 2nd try) No failures
Phase 3	Score of those passed: 98.70 (n=81) Score of those failed*: 75.0% (n=1)	Score of those passed: 96.73% (n=71) : 69 passed on 1st try; 2 passed on 2nd try) No failures
Overall for those completed training	98.02% (n=81)	96.51% (n=71)

*Participants were deemed to have failed in training when they failed to meet the 80% criterion after two attempts.

APPENDIX D
CONFEDERATE TRAINING GUIDE



Contents

Chapter 1: Generic Confederate Details

At the completion of this section, you should be able to:

- ✓ Understand what role a Confederate plays in scientific experimentation.
- ✓ Recognize an appropriate casting choice for performing the confederate role.
- ✓ Know and follow the general rules associated with being a confederate.

Introduction

To begin, you have been chosen as Confederate Experimenter to participate in one of three roles during a Nuclear Power Plant (NPP) simulation experiment. This guide will provide you with details for general Confederate Training and details specific to the NPP experimentation for each role. You should have previously read the Simulation User Guide and completed the Participant Training. You should also be well versed on how to navigate through the simulator and what procedures should be completed during the experiment (3-way communication, etc.). In order to prepare you for the Confederate roles, you must complete this training with minimal errors. Your performance will be periodically assessed during training. Additionally, ongoing evaluation will be administered throughout experimentation in order to ensure the highest standards of performance. The goal for this training is to receive consistent performance across all Confederates.

Confederate:

A Confederate is a Researcher performing the role of a Participant during the experiment. It is intended that the Confederate's status stays unknown to the Participant during the entire experiment. The confederate assists in manipulating the scenario in order to observe individual differences during team tasks.

Chapter 2: Nuclear Power Plant Specific Confederate Details

At the completion of this section, you should be able to:

- ✓ Recognize each Experimenter role, including your own.
- ✓ Understand the rules and details of your particular role.
- ✓ Know the procedures needed to be taken if complications arise.

Experimenter Roles

You will be assigned ONE duty, coinciding with specific roles carried out in a NPP. The specifics of each role are detailed below. Please review each role, but be sure to focus on characteristics of your specific role in depth.

Reactor Operator (**RO**): This job includes searching control panels to locate gauges, switches, lights, etc. The RO must accurately determine the state of a given indicator, communicate (using 3-way communication) the state to the other members of the team, and perform actions to change the state of the NPP when necessary.

Senior Reactor Operator (**SRO**): This job requires continual monitoring of RO stations and maintenance of 3-way communication. The SRO guides ROS through symptom based procedures to identify events/causes for system alarms.

Other Experimenter Roles Include the Instructor and the Lead Experimenter. These roles are briefly discussed here but further explanation is not necessary. In subsequent sections the roles of RO and SRO will be discussed in detail.

Instructor: The job of the instructor is to teach Participants about the NPP Simulator, guide them through the learning process, and conceal his or her familiarity with Confederate Experimenters.

Lead Experimenter: The job of the Lead Experimenter is to organize and monitor the progress of the experiment. This includes, script writing, scheduling, designing

Who Can Participate as a Confederate?

Generic guidelines must be followed in order to determine optimal Confederate selection. Confederates should be selected from the Participant population. If Undergraduate Students are used as the sample population, the Confederate should look like a plausible Undergraduate Student. In this case, Undergraduate Students can be of many different ages, races, etc. Thus, there are few limitations to the appearance of the Confederate casting. Confederates also should not be professional actors. Someone familiar with the experimentation process and scientific data collection practices should be used in order to ensure experimental methods are followed correctly. Furthermore, Confederates should not be acquainted with the Participant and vice versa. In order to reduce the risk of this happening during the experimental session, SONA Systems must be checked before each experimental session.

You have been chosen to become a Confederate because of your familiarity with scientific research methods and because you possess similar characteristics to the population being used in this experiment.

If for any reason during the study you feel you are acquainted with the Participant, please notify the Lead Experimenter on the project immediately. Sufficient time is needed to reschedule Confederates in this case.

Generic Rules for Confederates

Below is a list of rules for Confederate participation. These are necessary to the experimentation process and should be strictly followed:

1) **Arrive early** to each experimental session. Have your "SONA Participant number" ready. Additionally, have props such as purses, wallets, keys, etc. visible to the participant.
2) **Dress as the population**. This includes nice casual clothes. Do not wear business attire, but on the same note, do not wear unkempt, dirty clothes either. Demo days will be announced ahead of time; however, if you are scheduled to be a Confederate on a day that a demo is scheduled, please still dress in your Confederate attire. A few examples of appropriate attire would be jean pants and a polo shirt for men or jeans and a nice blouse for women.
3) **Do NOT ad lib**! This is especially important. In order to keep all variables consistent you need to stick to the script provided to

you upon the completion of this training. It is essential that you take the time to memorize this script word-for-word. Ample time will be given at work for you to practice this script; however, it is important to note that if you feel like you need additional practice, you should complete this on your own time.

4) **Keep your speech realistic and natural**. This will come with practice. Natural speech is needed in order to immerse the Participant into the experience and to ensure he/she remains unaware that you are an Experimenter rather than another Participant.

Chapter 2: Nuclear Power Plant Specific Confederate Details

At the completion of this section, you should be able to:

- ✓ Recognize each Experimenter role, including your own.
- ✓ Understand the rules and details of your particular role.
- ✓ Know the procedures needed to be taken if complications arise.

Experimenter Roles

You will be assigned ONE duty, coinciding with specific roles carried out in a NPP. The specifics of each role are detailed below. Please review each role, but be sure to focus on characteristics of your specific role in depth.

Reactor Operator (**RO**): This job includes searching control panels to locate gauges, switches, lights, etc. The RO must accurately determine the state of a given indicator, communicate (using 3-way communication) the state to the other members of the team, and perform actions to change the state of the NPP when necessary.

Senior Reactor Operator (**SRO**): This job requires continual monitoring of RO stations and maintenance of 3-way communication. The SRO guides ROS through symptom based procedures to identify events/causes for system alarms.

Other Experimenter Roles Include the Instructor and the Lead Experimenter. These roles are briefly discussed here but further explanation is not necessary. In subsequent sections the roles of RO and SRO will be discussed in detail.

Instructor: The job of the instructor is to teach Participants about the NPP Simulator, guide them through the learning process, and conceal his or her familiarity with Confederate Experimenters.

Lead Experimenter: The job of the Lead Experimenter is to organize and monitor the progress of the experiment. This includes, script writing, scheduling, designing

experimental scenarios, etc. **If there are any complications during the experiment please, bring issues to the Lead Experimenter immediately.**

Reactor Operator (RO)

Your objective as the RO Confederate is to participate in the activities mentioned above while also being able to understand and recognize contextual and scripted cues from both the Researchers and the Participants as to stay on task without priming or biasing the participant in any way.

There are 2 RO roles: RO1 and RO2. You will be assigned ONE of these roles. RO1 will be in charge of generating energy in the NPP, whereas RO2 will be in charge of all items in containment, including the reactor. Each role will have a different script. You will only have one script to memorize. If you need to trade your session for any reason please ensure your replacement has been trained to play the SAME role. Note that experimentation may only include the need for one RO Confederate Role.

Your tasks will include:

1) **Partake in the scenario acting as a Participant RO**. It will be necessary for you to portray unfamiliarity with the other Experimenters in the room. Following your script, you will be required to occasionally ask for clarification from other Experimenters, show confusion during specific task components, and to address other members by role not by name.
2) **Identify** contextual cues from Participants and scripted cues from Experimenters, while also being able to complete the experimental task. Contextual cues are unforeseen Participant reactions such as aggression, frustration, etc.
3) **Navigate** panels effectively. This is outlined in your Participant training. Be able to open panels, locate items of focus, and understand how to work simulator components thoroughly.
4) **Participate in 3-way communication** as instructed to do so in Participant training. Remember to address others by role and not name. State the action desired, determine the item of focus, identify the state of the item and confirm information given and received, but remember to do so without alerting the Participant to your familiarity with the tasking.

Your target behaviors, as explain above, should include showing unfamiliarity with the other Experimenters, slight confusion with the simulator interface, and some mild frustration with the scenario task. **Narrative details for these target**

behaviors will be scripted into the scenarios so make sure to study and memorize your script word-for-word.

Senior Reactor Operator (SRO)

Your objective as the SRO Confederate is to participate in the activities mentioned above while also guiding Participants through NPP scenarios. Additionally, you must recognize when ROs need help keeping Participants on task without alerting the Participants of the familiarity you have with other Experimenters.

Your tasks will include:

1) **Instruct ROs on which panels and instruments to locate**. This consists of knowing the location of all items of focus, understanding who to address and when, being able to correct RO errors (scripted and non-scripted), and correcting RO errors in a firm, but polite manner.
2) **Participate in 3-way communication** as instructed to do so in Participant training. Remember to address others by role and not name. State the action desired, determine the item of focus, and confirm information given and received, but remember to do so without alerting the Participant to your Experimenter status.
3) **Monitor Simulation and RO activity**. Identify where ROs are pointing to on the panels. Then determine if ROs are correctly completing steps given in the procedures. Continue to monitor the reactions of the participants and help alleviate any issues that may arise.

Target behaviors for the SRO role include unfamiliarity with RO Confederate and confidence in simulator navigation and materials. **Narrative details for these target behaviors will be scripted into the scenarios so make sure to study and memorize your script word-for-word.**

NPP Specific Rules for Confederates

This section will describe in detail the specifics of dialogue that can and cannot be said during experimentation, specifics of the script, and guidelines for how to deal with cancelations, etc.

Appropriate Word Choice

1) **Do not use NPP jargon!** Unless the terms have been used in the participant training guide, such as "RO" and "SRO", do not use acronyms and jargon terms. This is done to ensure that you do not alert participants as to how familiar you are with the Simulator and NPP processes.
2) **Do not address other members by name**. All members must be addressed by role not name.
3) **Do not refer to past or future performance.** Be aware of what you say and how you say it. For example, do not say "this is how I did it last time." Always remain focused and present in your current session! Do not get your sessions mixed up.

Script Details

Script narrative directions are a key element during experimentation. The consistency of delivering these lines between Experimenters is incredibly important to the outcome of the research.

Some notes to remember are:

1) **Be aware of "stage" directions**. Make sure to integrate any of these instructions into your acting.
2) **Allow for ample time to practice the script**. It needs to be word-for-word. Do not procrastinate. Multiple evaluations will be given before experimentation begins in order to determine your ability to play the role given to you.

If Participants cannot run for any reason and need to reschedule, note who the RO Confederate was in order to ensure he/she does not run with that particular Participant again. Additionally, remember to provide the debriefing form and the evaluation form at the end of each session, this is in congruence with IRB rules and regulations. At the very end of the session the participant should know that you are a confederate, thus there is no need to act out "leaving" the session.

Chapter 3: Performance Evaluation

At the completion of this section, you should be able to:

- ✓ Understand what is expected of you before during and after experimental sessions.
- ✓ Know and follow the guidelines for filling out evaluation forms.

Evaluation Methods

You will be required to complete chapter quizzes during training and 3 training sessions in which you will go through the script with other Confederates. Furthermore, you will also be asked to fill out After Session Reports during experimentation after each session has completed.

Chapter Quizzes

You will be given chapter quizzes during the Confederate Training presentation; each section with have a corresponding evaluation. Please be sure to read questions carefully. You are only allowed to miss one answer on each of these in order to pass on to the next section of training.

Narrative/Script Training

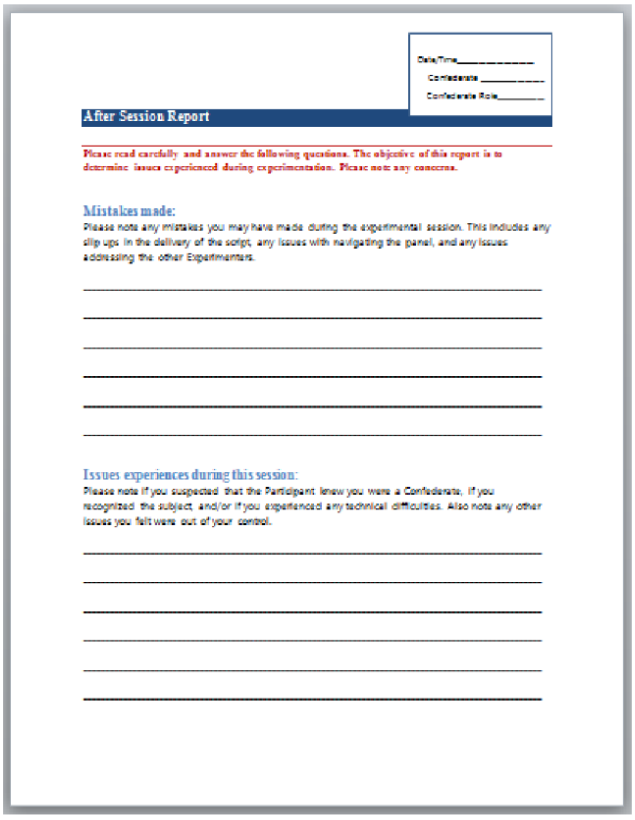
These sessions will be recorded for your convenience. In these sessions Confederates will work together to refine narrative details and provide any feedback to the Experimenters about any issues with the script. Additionally, you are encouraged to playback your performance in order to improve. These sessions will be completed before experimentation begins.

After Session Reports

After each experimental session is finished, you will be asked to complete an After Session Report. In these reports, you will be noting any mistakes you made during the experiment, impressions of how the session went, if you suspected that the Participant knew you were a Confederate, if you

recognized the subject, and/or if you experienced any technical difficulties. These reports will be evaluated against the reports of the other Experimenters and the Experimenter logs.

This is an example of the After Session Report



Date/Time_____
Confederate _____
Confederate Role_____

After Session Report

Please read carefully and answer the following questions. The objective of this report is to determine issues experienced during experimentation. Please note any concerns.

Mistakes made:

Please note any mistakes you may have made during the experimental session. This includes any slip ups in the delivery of the script, any issues with navigating the panel, and any issues addressing the other Experimenters.

Issues experiences during this session:

Please note if you suspected that the Participant knew you were a Confederate, if you recognized the subject, and/or if you experienced any technical difficulties. Also note any other issues you felt were out of your control.

Chapter 4: Scheduling

At the completion of this section, you should be able to:

- ✓ Understand what time commitments are required of you for training and experimentation.

Schedule Requirements

During the experimentation development and implementation, you will be asked to attend frequent training sessions, meetings, and experimental sessions.

As discussed above, you will be asked to complete 3 training sessions in order to practice the narrative provided. Choose one or two partners (depending on the experimental need) to practice with who have been assigned a different role than yours. Upon the completion of this training, you and your partner(s) must decide on 3 timeslots in which you will complete your videotaped script evaluation.

Throughout the experimentation you should be aware that complications may arise with scheduling. **If for any reason you cannot complete a session, let your Lead Experimenter know immediately**. Also, if someone cannot complete their scheduled session, please be considerate and trade sessions with them if your personal schedule allows it.

2022-11

HUMAN PERFORMANCE TEST FACILITY (HPTF)

VOLUME 1 - Systematic Human Performance Data Collection Using Nuclear Power Plant Simulator: A Methodology

2022-11 HUMAN PERFORMANCE TEST FACILITY (HPTF)