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

RIL 2022-11

HUMAN PERFORMANCE TEST FACILITY (HPTF)

VOLUME 3 - SUPPLEMENTAL EXPLORATORY ANALYSES OF SENSITIVITY OF WORKLOAD MEASURES

Date Published: January, 2023

Prepared by:
N. Hughes[1]
J. Lin[2]
G. Matthews[2]
D. Barber[2]
K. Dickerson[1]

U.S. Nuclear Regulatory Commission
Office of Nuclear Regulatory Research
Division of Risk Assessment
Human Factors and Reliability Branch
Washington, DC 20555-0001

Institute for Simulation and Training, University of Central Florida
3100 Technology Pkwy
Orlando, FL 32826

Niav Hughes & Kelly Dickerson 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 staff of the U.S. Nuclear Regulatory Commission (NRC) is responsible for reviewing and determining the acceptability of new reactor designs and modifications to operating plants to ensure they support safe plant operations. Human factors (HF) staff use Chapter 18 of the Standard Review plan (NUREG-0800, NRC, 2007) and the guidance documents referenced therein, in part, to ensure that plant operators can safely control the plant. The NRC's Human Factors Engineering (HFE) Program Review Model, NUREG-0711, Rev. 3 (NRC, 2012) is one of these documents. NUREG-0711 outlines that a generic "human centered" HFE design goal should include a design that supports personnel in maintaining vigilance over plant operations and provide acceptable workload levels. Furthermore, NUREG-0711's review elements highlight the importance of considering workload (WL), particularly, in the review criteria for Elements 5 (Task Analysis) and 6 (Staffing and Qualifications). Elements 5 and 6 indicate explicitly that an estimate of WL must be part of the review of the HFE design in order for the reviewer to make a determination of reasonable assurance of safety.

The basis for current NRC HFE guidance, comes (in part) from data based on research conducted in other domains (e.g., aviation, defense), qualitative operational experience in nuclear power plants (NPPs), and a limited number of empirical studies in a nuclear environment. When it comes to new designs, technologies, and concepts of operations for new control rooms, there may not be operational experience and appropriate research literature to draw from to inform NRC HFE guidance. To address this gap in the research, the Commission, in Staff Requirements Memorandum (SRM) SECY-08-0195, directed the staff to consider using generic simulator platforms to address human performance issues. In response to the SRM, the Office of Nuclear Regulatory Research (RES) developed the NRC Human Performance Test Facility (HPTF) research program. The HPTF empirically measures and studies human performance aspects of control room operations using a NPP simulator and a combination of objective and subjective measures of workload. The information gained will be utilized to enhance the technical basis for the NRC's regulatory guidance in HFE and to better inform models for human reliability analysis (HRA).

To date, four large-scale data collections have been performed for the HPTF research program. The experiments are reported in RIL 2022-11 Volumes 1 and 2. In order to delve further into the data previously collected, we performed a reanalysis of all HPTF experiments thus far to further 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. The present RIL 2022-11 Volume 3 describes the supplementary analyses performed on datasets from four HPTF experiments to further investigate workload measure sensitivities.

Previous HPTF reports have documented how physiological and subjective workload measures are diagnostic for assessing the impacts of task type, simulator type and operator expertise (Reinerman-Jones & Mercado, 2014; Reinerman-Jones et al., 2016, 2018, 2019). However, diagnosticity does not ensure sufficient sensitivity. In fact, previous studies using multiple measures of workload reveal a range of different sensitives across the different task types. In addition, this research is novel given that few studies have directly compared the various objective and subjective workload metrics for their sensitivity in the nuclear power plant operation domain.

Understanding which indices are more and less sensitive for measuring WL on the performance of control room tasks will provide useful insights to human factors licensing technical staff for their assessment of an applicant's HFE program (NUREG-0711) and staffing analyses (NUREG-1791). For instance, to demonstrate successful implementation for some of the elements in their HFE design program using the guidance found in NUREG-0711, applicants propose a variety of metrics to measure workload. Most often, the NASA-TLX, a subjective measure, is used, but there have been instances where applicants or licensees deviate from this precedent. Having a better understanding about which measures are more or less sensitive, in what context, and in comparison, to the NASA-TLX will enhance staff knowledge base in the use of these human performance metrics and aid technical reviewers' determination as to an applicant's correct use of the metric(s) chosen.

Analyses of effect sizes were used to characterize the magnitude of response or rating changes in the workload metrics. In addition, this report provides a summary of the convergence and divergence of the NASA-TLX with other workload measures. Overall, the analyses suggest that many of the workload measures utilized in the HPTF studies show practically relevant sensitivity to the workload changes induced by the experimental manipulations in the simulated NPP operations. These results indicate reasonable confidence that the measures used discern meaningful differences in terms of WL for control room tasks in a variety of contexts (e.g., novice vs former operator; interface technology, partial scale versus full scale simulator).

Convergence of NASA-TLX and some psychophysiological measures provides good indication that NASA-TLX, which is the most commonly used subjective measure is sufficiently sensitive in practice and provides reliable estimates of operator workload. The most commonly used physiological metric is heart rate, however, the sensitivity profile of this measure was less straightforward. For example, interpreting cardiac data needs to be done with caution, especially when verbal communication and physical movement is involved. Taken together the results of these analyses demonstrate good consistency among the measures such that technical review staff can have reasonable confidence in each of the measures analyzed. Overall, considering the complexity of NPP operations and workload variation involved, using multivariate assessment of workload facilitates a more comprehensive understanding of impacting factors of workload.

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. Additionally, the empirical human performance data collection ensures a better understanding of the various cognitive and physical elements that support safe control room operation.

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.

The findings from the resulting experiments are presented in a series of volumes, Research Information Letter (RIL) report, “Human Performance Test Facility (HPTF) (RIL 2022-11). Volume 1, titled “Systematic Human Performance Data Collection Using Nuclear Power Plant Simulator: A Methodology” contains two studies and compares performance, physiological, and subjective measures of workload in operators and novices in a simulated digital representation of an analog plant in both a touchscreen and desktop configuration. Volume 2, titled “Comparing Operator Workload and Performance Between Digitized and Analog Simulated Environments” contains a single study and compares formerly licensed operators’ performance, physiological, and subjective workload between a full scale, full scope simulator, and the HPTF’s lightweight digitized simulator environment during an emergency operating procedure scenario.

The present report, Volume 3, titled “Supplemental Exploratory Analyses of Sensitivity of Workload Measures” contains a re-analysis of data from volume 1 and 2 with the specific goal of determining sensitivity to workload variations for each of the subjective

and physiological measures. Understanding which indices are more and less sensitive to variations in workload will provide useful insights to HF licensing technical staff for their assessment of an applicant's HFE program (NUREG-0711) and staffing analyses (NUREG-1791). For instance, applicants or licensees typically only use subjective measures of workload to demonstrate successful implementation of some elements of their HFE design program. The NASA-TLX is the most used assessment technique, however, there have been instances of deviation from this precedent. Having a better understanding about which NASA-TLX alternatives are more or less sensitive, and in what context, will aid technical reviewers' determination as to an applicant's correct use of the metric(s) chosen. Additionally, studies like Volume 3 aid in understanding the underlying physiological mechanisms that drive the changes in self-assessed workload using subjective measures alone.

Chapter 1 of this report begins with a description of workload assessment for NPP operations, presents background on the HPTF research program, and outlines the aim of the study to further investigate the sensitivity of workload measures for the data reported in RIL 2022-11 Volumes 1 and 2. Chapter 2 provides the methodological approach and methods employed as well as a a summary of each of the previous studies. Chapter 3 reports the results using effect sizes to quantify the measure sensitivity. Chapter 4 contains the discussion of the findings that workload (WL) responses induced by task type manipulations vary in magnitude, depending also on sample and interface type, suggesting sensitivity variation of the metrics in detecting WL changes depending on the assessment circumstances. Chapter 5 describes the significance of understanding sensitivity of WL measure used in the nuclear domain including implications for HFE guidance development and enhancements to HRA models.

1 INTRODUCTION

1.1 Workload Assessment Methodology for Nuclear Power Plant Operations

The Human Factors Engineering (HFE) staff of the Nuclear Regulatory Commission (NRC) evaluate the HFE programs submitted in license applications for nuclear power plants (NPPs) to ensure their safety. One element of the review is to determine appropriate function allocation which is the allocation of functions between operators and automatic control systems which are then separated into tasks. “Function allocation is the assignment of functions to (1) personnel (e.g., manual control), (2) automatic systems, and (3) combinations of both. Exploiting the strengths of personnel and system elements enhances the plant’s safety and reliability, including improvements achievable through assigning control to these elements with overlapping and redundant responsibilities. Functions are allocated to human and system resources and are separated into tasks. The subsequent analysis of personnel tasks identifies the alarms, displays, controls and task support needs required for performing the task. Tasks are arranged into jobs and assigned to staff positions or roles within the control room (e.g., reactor operator, balance of plant). Each position is evaluated to verify the workload (WL) is acceptable.” (NUREG-0711, NRC 2012). As such, due consideration should be given to whether there are aspects of operator tasking that are liable to impose excessive cognitive WL and so raise error probabilities and threats to safety. WL assessment can contribute to prospective Human Reliability Analysis (HRA) for the nuclear industry (NUREG/CR-1278, NRC, 1983). HRA seeks to model and identify potential contributors to human error, with the ultimate aim of quantifying error likelihoods (Boring, 2012). In the NPP context, HRA may be especially valuable as an approach to assessing and minimizing risk in next generation control rooms (Tran, Boring, Joe & Griffith, 2007).

The value of WL assessment for HRA is that it can help to identify relevant performance shaping factors that raise error probabilities, especially those derived from task demands. While HRA traditionally focuses on predicting error rates on a probabilistic basis, contemporary approaches aim also to model the cognitive processes that underlie human performance NUREG-2198 (NRC, 2020); RIL 2020-02 (NRC, 2020); NUREG-2114 (NRC, 2012); Mosleh & Chang, 2004). WL assessment contributes to quantifying these processes and their sensitivity to task demands. In addition, factors influencing performance are often dynamic and interdependent, and continuous psychophysiological monitoring of operator state provides a means for tracking performance-influencing factors dynamically (Tran et al., 2007).

WL assessment is especially important in the NPP context because reactors in the United States utilize a variety of plant designs, interfaces, and safety systems. For example, the main control room (MCR) must be designed differently depending on whether the plant is a boiling water reactor (BWR) or a pressurized water reactor (PWR), and the WL factors in the two types of plant may thus differ. In addition, MCR designs are evolving to reflect plant modernization. For example, there may be impacts on task demands from new interface features such as touchscreens. Moreover, how, and where these new interfaces are implemented may additionally impact task demands and operation. The diversity of designs requires a standard WL assessment methodology that can in turn support a systematic HRA process.

This report describes analyses to investigate the sensitivity of WL measures (Lin et al., 2021) in simulated NPP MCR operations, supplementing previous reports addressing factors such as task type, interface type, and operator experience (Reinerman-Jones & Mercado, 2014;

Reinerman-Jones, Teo & Harris, 2016; Reinerman-Jones et al., 2018, 2019; RIL 2022-11 Volume 1, in press). This introduction provides a summary of the work performed as detailed in existing reports, and the motivation for performing the additional analyses.

1.2 The Human Performance Test Facility Research Program

The program of research known as the NRC Human Performance Test Facility (HPTF) has aimed to support the NRC's mission by advancing, validating, and documenting WL assessment methodology for NPP MCR operations using a generic plant simulator (Hughes, D'Agostino, & Reinerman-Jones, 2017). Using these simulators, the Human Factors and Reliability Branch (HFRB) in the Office of Nuclear Regulatory Research (RES) began a program of research known as the NRC HPTF to collect empirical human performance data with the purpose of measuring and ultimately better understanding more about the various cognitive and physical elements that support safe control room operation. In order to leverage expertise in robust experimental design as well as access to a large sample population (i.e., university students), the NRC partnered with a university. The HFRB staff worked as co-investigators along with a team of researchers at the University of Central Florida (UCF) Institute for Simulation and Training (IST) to design and carry out a series of experiments aimed at measuring and understanding the human performance aspects of common control room tasks through the use of a variety of physiological and self-report metrics.

Controlled experimental studies of WL response conducted at UCF and NRC headquarters and technical training center locations have utilized the HPTF methodology constructed with support from NRC, including inputs from human factors and nuclear operations Subject Matter Experts (SMEs). The HPTF provides a facility for assessment of the impact of novel designs, technologies and concept of operations on operator WL and performance using human-in-the-loop experiments. It is centered on a GSE Generic Pressurized Water Reactor (GPWR) simulator that can be configured to provide experimental control over the task elements performed by operators. The GSE GPWR simulator has the capability to be full-scope and is adapted for simulating specific experimental scenarios as a part-task simulator. It is still intended that the simulator will produce results which are generalizable to full-scope simulators. Experiments have used a modified generic Emergency Operating Procedure (EOP) that requires participants to perform predetermined tasks to respond to a loss of all alternating current power to the plant's safety buses (EOP-EPP-001 GSE Power Systems, 2011). Four key features of the methodology are the use of NPP MCR simulated environments, novice participants, the definition of task components, and multivariate WL assessment using both subjective and objective measures.

1.2.1 Types of NPP simulated environments

The use of a real NPP simulator to create a realistic experimental environment is a cornerstone of the HPTF methodology. As NPP reactor technology and control room design has modernized and evolved, so too has the NPP simulator technology and capability. In the HPTF studies, we characterize the types of NPP simulators with five main features, summarized in Table 1.

Table 1. Summary of types of NPP simulated environments.

Features	NPP Simulator Types
Scope	a. Full scope simulator – has the capability to simulate all of 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 – 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 – 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	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)

Based on these definitions, the simulators used in the HPTF studies can be characterized into three types: 1. a full-scope simulator with hierarchical layout, simulated analog interface in sit-down desktop mouse click workstations; 2. a full-scope simulator with hierarchical layout, simulated analog interface in stand-up touchscreen workstations; 3. a full-scope simulator with spatially dedicated layout, analog interface in stand-up manual manipulation benchboard.

1.2.2 Use of novice and experienced participants

Control room operations require a team that includes Reactor Operators (ROs) and a Senior Reactor Operator (SRO). Typically, in traditional control room operations, the SRO orchestrates the progress of plant operations, initiates three-way communication procedures crucial to successful task completion, and provides the ROs task instructions when necessary. Operators represent expert performers, but, for research purposes, it may be challenging to recruit licensed operators for research, especially given the need for multiple crew members. To address this practical limitation this program of research has investigated and found support for the use of novice participants, defined as those without industry experience. Tasks used for experimentation are designed to minimize the role of prior experience and knowledge, but still impose cognitive demands on the critical elements of information-processing for real performance. These include working memory, selective and sustained attention, and manual response selection and execution. From a cognitive engineering standpoint, experimental studies can reveal processing operations that may be vulnerable to overload in novices and experts alike. That is, an “equal but different approach” is taken to ensure that cognitive demands are comparable across populations, but the knowledge requirements are calibrated to the skill-base of novice participants. Similarly, the physical environment can be simplified for novices by reducing the number of controls within each display panel and simplifying the naming convention of specific gauges and switches (Reinerman-Jones, Guznov, Mercado, & D’Agostino, 2013).

1.2.3 Operator task classification

The experimental scenario used for the studies described here is a modified version of ECA-0.0, Loss of All Alternating Current Power. The EOP represented in the simulation was decomposed into a series of discrete tasks labeled checking, detection, and response implementation, which can be readily trained within the novice population. These tasks are representative of tasks performed primarily by ROs and directed by SROs (NUREG/CR-6947, 2008; O'Hara & Higgins, 2010; Reinerman-Jones et al., 2013). *Checking* requires a one-time inspection of an instrument or control to verify that it is in the appropriate state. *Detection* requires continuous monitoring of a control parameter to identify a change in the state of the plant. *Response implementation* requires a fine motor response (mouse usage or finger touch) to change the state of the NPP by locating a control and subsequently manipulating the control in the required direction. The experimental protocol represents the EOP as a sequence of steps using these three types. The temporal order of tasks can be manipulated but, in an actual NPP EOP, checking always precedes response implementation, while detection can occur at any point. Thus, possible task type sequences include: (1) checking, response implementation, and detection, (2) checking, detection, and response implementation, and (3) detection, checking, and response implementation.

1.2.4 Multivariate workload assessment

There has been a longstanding debate in human factors over the optimal methodology for workload (WL) assessment. A major challenge has been that different measures may dissociate (Hancock & Matthews, 2019). That is, manipulations of task demands may have different impacts on subjective WL, psychophysiological indicators of brain response, and objective performance metrics. Thus, while the NASA Task Load Index (NASA-TLX: Hart & Staveland, 1988) is the single most popular WL measure, it does not provide a comprehensive WL assessment. Indeed, relying on the NASA-TLX may lead to neglect of task factors whose impacts on WL, and hence error probability, require psychophysiological measures.

Work conducted in the HPTF has supplemented the NASA-TLX with additional subjective measures including the Multiple Resource Questionnaire (MRQ: Boles & Adair, 2001) which has greater diagnosticity for different sources of demand such as working memory and spatial attention. Stress response is assessed with the Dundee Stress State Questionnaire (DSSQ: Matthews et al., 2002). WL is also assessed with an integrated suite of psychophysiological sensors, summarized in Table 2. Performance measures include those capturing effectiveness of three-way communication as well as those that index accuracy of task execution. Taken together, these multiple subjective and objective measures provide a comprehensive picture of operator response to changing task demands.

Table 2. Summary of sensors and metrics used for WL assessment at the HPTF

Sensor	Method	Metrics
Electrocardiogram (ECG)	Typical electrode placement: single-lead electrodes on the center of right clavicle and lowest left rib	Heart rate (HR), Inter-beat interval (IBI), Heart rate variability (HRV)
Electroencephalogram (EEG)	Multiple scalp electrodes at frontal, temporal, parietal, and occipital sites	Spectral power densities (SPDs) for frequency bands (delta, theta, alpha, beta)
Cerebral blood flow velocity (CBFV)	Transcranial Doppler (TCD) ultra-sonography using transceivers above zygomatic arch	Bilateral CBFV in middle cerebral arteries Task-induced response

Functional near-infrared spectroscopy (fNIRS)	Forehead IR light sources and detectors to measure prefrontal blood oxygenation	Bilateral cortical oxygenation in the prefrontal cortex

Findings from the HPTF studies have been summarized in a series of reports and articles (Reinerman-Jones & Mercado, 2014; Reinerman-Jones et al., 2016, 2018, 2019). A full summary of the many findings of the previous studies is beyond the scope of the present report. One consistent theme that is of central focus for this report is that the level of WL measured across multiple metrics was different for each of the three task types. Specifically, there was a convergence of higher WL ratings and objective measurements for the detection task, which is consistent with the human factors research on vigilance and sustained attention (Warm, Parasuraman & Matthews, 2008), identifying this element of tasking as a potential vulnerability and a focus for HRA. Studies using novice samples have provided the statistical power required to define WL responses accurately. Additional studies have confirmed that WL factors generalize to experienced populations assuring that findings are relevant to operational practice. Experienced participants include both well practiced HPTF researchers (Leis et al., 2014) and former NPP operators tested at a simulator at NRC headquarters (Reinerman-Jones et al., 2018) and at the NRC Technical Training Center in Chattanooga, Tennessee (Reinerman-Jones et al., 2019).

1.2.5 Sensitivity of workload measures

Results from assessment of workload (WL) using multivariate strategies suggested that the measures of WL may differ in their sensitivity to task types. In addition, few studies have directly compared the various objective and subjective WL metrics for their sensitivity in the nuclear power plant operation domain. It is worth diving more deeply into the data collected in previous experiments to compare the sensitivity of the various WL indices to task type in terms of task distribution (i.e., blocks of tasks versus full scenario). Understanding which indices are more and less sensitive for measuring WL on the performance of control room tasks will provide useful insights to human factors licensing technical staff for their assessment of an applicant's HFE program. For instance, to demonstrate successful implementation for some of the elements in their HFE design program using the guidance found in NUREG-0711, applicants propose a variety of WL metrics. Most often, the NASA-TLX, a subjective measure, is used, but there have been instances where applicants or licensees deviate from this precedent. Having a better understanding about which measures are more or less sensitive, in what context, and in comparison, to the NASA-TLX will enhance staff knowledge base in the use of these human performance metrics and aid technical reviewers' determination as to an applicant's correct use of the metric(s) chosen.

In order to investigate the sensitivity of measures, we chose to use effect sizes to quantify the magnitude of response or rating changes in the WL metrics. By comparing the effect sizes of each metrics in different task type manipulations, we can identify which metric is more sensitive in picking up the changes in responses to WL induced by a specific task type. We can also determine whether certain metrics are more sensitive across the board, or whether differences in metric sensitivity depend on the interface and participant experience. Given that the industry typically utilizes the NASA-TLX (Hart & Staveland, 1988) to evaluate WL, a particular concern is whether the NASA-TLX is sufficiently sensitive to pick up vulnerability to overload across a range of interfaces and operator characteristics, or whether there are contexts in which psychophysiological metrics can pick up vulnerabilities which might not be apparent in NASA-TLX data.

1.3 Aims

The aim of the research reported here was to analyze data from previous HPTF studies to further investigate the sensitivity of the WL measures employed. Data from four studies were utilized for this purpose, two using novice samples (Studies 1 and 2), and two using former operator samples (Studies 3 and 4). Studies 1-3 used a common scenario based on the modified ECA-0.0 for loss of all alternating current power executed using the GSE Generic PWR simulator. The scenario was presented using a digital, part-task simulator which allowed the number of each of the three task types to be equated using a blocking method for experimental control purposes. There were three different orderings used, allowing tests for the impacts of certain task orders. Study 4 was conducted in a full-scale, full-scope simulator environment that reproduced both the physical environment and the would-be physics of a real plant and plant response. It used a more realistic but also more complex sequence of steps and execution of a full scenario for ECA-0.0.

2 SUMMARY OF METHODS

The HFE staff of the NRC evaluates the HFE programs of applicants for construction permits, operating licenses, standard design certifications, combined licenses, and amendments to licenses. The purpose of these reviews is to support public health and safety. NRC's *Human Factors Engineering Review Model* (NUREG-0711) provided a "top-down" approach to conducting HFE program safety evaluation. According to this guidance, a review should start at the "top" with an overview of the high-level plant goals and then define the functions necessary to achieve the goals. Functions are allocated to human and system resources and subsequently separated into tasks for specifying the alarms, information, controls, and task support needs needed to complete functional assignments. Tasks are arranged into jobs and assigned to staff positions. Each position should be evaluated to verify the WL for the assigned tasks is acceptable (NRC, 2012).

To verify that the WL is acceptable, WL must be defined and properly assessed. According to NUREG-0711 (NRC, 2012), WL is comprised of the physical, cognitive, and other demands that tasks place on plant personnel. However, the guidance did not specify what method should be used to assess WL in the NPP domain in different tasks and scenarios. A multivariate strategy was used in the HPTF experiments. Full details of the methods for these studies are provided in the reports already delivered to the NRC (Reinerman-Jones & Mercado, 2014; Reinerman-Jones et al., 2015, 2018, 2019). Here, we provide an overview only, with a focus on investigating the sensitivity of WL measures in the NPP domain.

The multivariate assessment of WL used in the HPTF included both subjective measures (self-assessment rating scales) and objective measures (psychophysiological and performance-based measures). This report focuses on investigating and comparing the sensitivity of the subjective measures and psychophysiological measures.

2.1 Subjective Measures

2.1.1 NASA-Task Load Index (NASA-TLX)

The NASA-TLX (Hart & Staveland, 1988) is a widely used multi-dimensional measurement of subjective WL. In the HPTF studies it was used to measure the perceived WL at the end of each

task type or after the entire experimental scenario, depending on the study. It consists of six separate rating scales for workload-relevant factors:

- mental demand
- physical demand
- temporal demand
- performance
- effort
- frustration.

All factors, except performance, are rated on a 0 - 100 scale from “Low” to “High”. Performance is rated on a 0 - 100 scale from “Good” to “Poor”.

2.1.2 Multiple Resource Questionnaire (MRQ)

The MRQ was used to characterize the nature 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 (Boles, 1991, 1992, 1996, 2002). Participants received a copy of the scale with definitions and completed the MRQ at the end of each task type or the scenario depending on the study using a computerized version of the questionnaire. The MRQ methods suggest using only the task relevant scales. The following 14 of 17 scales were included for the present study which can be roughly grouped into 5 subscale categories, these are: language related, visual, spatial, action, and general cognitive (see table 3).

Table 3. MRQ Subscale Categories.

Language Related Subscales	**Visual Subscales**	**Spatial Subscales**	**Action Related Subscales**	**General Cognitive Subscales**
Auditory emotional process	Visual lexical process	Spatial attentive process	Manual process	Short term memory process
Auditory linguistic process	Visual phonetic process	Spatial concentrative process		
Vocal process	Visual temporal process	Spatial emergent process Spatial positional process Spatial quantitative process		

2.1.3 Instantaneous Self-Assessment (ISA)

The ISA (Tattersall & Foord, 1996) is a subjective unidimensional WL rating method that provides a continuous and concurrent assessment of task demand on perceived WL. In the HPTF studies, it was used to measure the perceived WL for each task type. Participants were asked to verbally rate their WL for completing each of the three task types (checking, detection, and response implementation) using a 5-point Likert scale ranging from “1 = Very Low” to “5 = Very High”.

2.2 Physiological Measures

2.2.1 Electroencephalogram (EEG)

The Advanced Brain Monitoring B-Alert X10 system uses nine-channels of EEG and one channel of ECG. Electrodes were placed following the international standard 10-20 System. Data were collected using a sampling rate of 256 Hz and signals were captures from Fz, F3, F4, Cz, C3, C4, Pz, P3, and P4 electrode sites. Reference electrodes were placed on each of the participant's mastoid bones. Power Spectrum Density (PSD) analysis was used to assess three standard bandwidths: theta (4-8 Hz), alpha (9-13 Hz), and beta (14-30 Hz) (Wilson, 2002). Each bandwidth was collected at each of the nine electrode sites. Data were then aggregated to compare left and right hemispheres and the front, temporal, and parietal lobes.

2.2.2 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. Heart rate was computed using peak cardiac activity to measure the interval from each beat per second. The "So and Chan" QRS detection method was used to calculate Inter-beat Interval (IBI) and Heart Rate Variability (HRV: Taylor, Reinerman-Jones, Cosenzo, & Nicholson, 2010). This approach maximizes the amplitude of the R-wave (Henelius, Hirvonen, Holm, Korpela, & Muller, 2009).

2.2.3 Transcranial Doppler (TCD)

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

2.2.4 Functional Near Infrared Imaging (fNIRS)

The Somantics' Invos Cerebral/Somatic Oximeter, model 5100C, was used to monitor (hemodynamic) changes in oxygenated hemoglobin and deoxygenated hemoglobin in the left and right hemisphere prefrontal cortex (Ayaz et al., 2011; Chance, Zhuang, UnAh, Alter, & Lipton, 1993).

2.3 Summary of Studies

The analysis to investigate the sensitivity of WL measures was based on reanalyzing the data collected in four previously completed studies in the NRC HPTF project. These four studies evaluated WL responses of participants from different populations (i.e., university students and former operators) using simulators with different types of interface (i.e., simulated analog with desktop interface, simulated analog with touchscreen interface, and analog interface) in three common NPP operation task types (i.e., checking, detection, and response implementation). Designs of the four studies are summarized in Table 4. More detailed information is provided in the following sections.

Table 4. Study Design Summary.

Study	Participant	Sample size	Role	Simulator	Interface

Study 1	Student novice	81	RO1	GSE GPWR	Desktop digital
Study 2	Student novice	71	RO1	GSE GPWR	Touchscreen digital
Study 3	Former operator	18	RO1/2	GSE GPWR	Touchscreen digital
Study 4	Former operator	30	RO1/2	TTC Westinghouse PWR	Analog

2.3.1 Study 1

This study aimed to confirm the feasibility of a novel methodology in the nuclear domain. It used novice participants to perform common NPP operator tasks in a simplified desktop-based simulated environment. Stimuli were presented on two 24-inch (16:10 aspect ratio) UXGA monitors. Participants used a mouse and scroll-wheel to view all the controls as not all the controls could fit in the display area of the monitors. Task performance required mouse and keyboard inputs. Participants were 81 UCF students (45 males, 36 females, *M* = 21, *SD* = 4.11) trained to an acceptable level of proficiency prior to the main WL assessment in the simulated EOP. The study confirmed that, out of the three tasks, the detection task imposed the highest WL on multiple metrics, evident in both subjective and objective metrics, including higher NASA-TLX scores, spatial-attentive and temporal WL, higher regional brain oxygenation (measured by fNIRS), and less accurate communication performance. Some specific WL indices showed differing trends, but in general the convergence between WL and performance data confirmed the HRA relevance of the assessment.

2.3.2 Study 2

The second study used a similar design, with the aims of testing generalization of findings to a touchscreen interface, and of identifying differences in WL and performance between desktop and touchscreen interfaces. The touchscreen interface consisted of eight 27-inch touchscreen WQHD (Wide Quad High Definition) monitor grids (two high by four wide). The interface displayed the instrumentation and control panel in its 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. Seventy-one participants (40 males, 31 females, *M* = 20.15, *SD* = 2.65) from the UCF student pool participated. The study confirmed that task type influenced multiple WL metrics and performance when using a touchscreen interface. Task type effects were comparable to those found for the desktop interface, with the detection task tending to produce the highest WL response, across multiple metrics. Some generally minor differences in detail in task type effects were found. Results also showed some differences in WL imposed by the two interfaces, depending on the metric examined. Findings served as an initial elucidation of some of the costs and benefits of introduction of touchscreens to the MCR as part of modernization efforts.

2.3.3 Study 3

The third study used a similar experimental design as studies 1 and 2, but rather than novices the participants were a sample of formerly licensed operator (*N*=18; 14 males, 4 females, *M* = 45.94, *SD* =10.63). Participants had operational experience working in a PWR or BWR MCR in either commercial power generation or naval nuclear power generation domains. The study aimed to determine whether comparable WL findings would be obtained from an expert sample, relative to the results from novice samples in Studies 1 and 2. Experimental sessions were

conducted in a mock MCR at the NRC headquarters in Rockville, Maryland. A GPWR NPP MCR simulator was configured for a crew of three operators, including an SRO and two ROs. A touchscreen interface was used comprising four 27-inch touch monitors arranged two high by two wide. It also distinguished former operators performing in RO1 and RO2 roles. Findings showed task type differences are broadly comparable to those demonstrated in the first two studies, suggestive of highest WL on the detection task. The study also demonstrated WL differences between RO1 and RO2 on some metrics.

2.3.4 Study 4

The fourth study aimed to validate the feasibility of the HPTF methods, procedures, and the *different but equal* paradigm and confirm the generalizability of the results from the first three studies using an analog, full-scope, full-scale[2] simulator at the NRC Technical Training Center (TTC) in Chattanooga, Tennessee, which replicates a Westinghouse 4-Loop Pressurized Water Reactor (PWR) design and has the capability to simulate all the physical and underlying thermodynamics occurring in the real plant. Former operators provided the sample (N = 30 males, M = 55.47, SD =7.82). Similar to Study 3, Study 4 distinguished former operators performing in RO roles. In this case, roles were designated as RO and BOP (Balance of Plant). By contrast with Studies 1-3, the scenario followed a realistic EOP, without attempting to experimentally control the frequencies and orders of the different task types. Overall, the study confirmed the generalizability of the previous findings and supported the feasibility of utilizing digital simulators to conduct research, identify safety concerns, and supplement operator training.

2.4 Experimental Scenario

Experimental scenario study 1-3

The experimental scenario consisted of tasks reflecting common activities required when completing operating procedures: checking (C), detection (D) and response implementation (R). Tasks were composed of individual steps, e.g., one checking operation. There were twelve steps in the experimental scenario, grouped by task type (4 checking steps, 4 detection steps, and 4 response implementation steps). The order of task type block was counterbalanced across participants. The task types were only partially counterbalanced to create scenarios because the tasks of checking and response implementation are directly linked such that checking always occurs before response implementation in real NPP operations. Task order must thus be constrained to maintain external validity.

Experimental scenario study 4

The experimental scenario was developed based on a generic version of an emergency operating procedure (EOP) for a “Loss of All Alternating Current (AC) Power (ECA-0.0)” scenario but modified for experimental use. The experimental procedure contained 69 steps supporting three different task types, i.e., checking, detection, and response implementation task types. In the experimental procedure, there were 30 steps (16 checking, 5 detection, and 9

[2] Full Scale simulator [at the TTC] refers to the control room layout which is spatially dedicated and continuously visible. In this layout arrangement, all the Instrumentation and Controls (I&C) are available to the operator and presented in a fixed location (e.g., existing fleet of Light Water Reactors, Westinghouse PWR).

response implementation) for RO, and 39 steps (27 checking, 1 detection, and 11 response implementation) for BOP. The number of steps was not balanced for RO nor was task type due to the nature of the original, realistic EOP, which requires steps to be taken in a prescribed sequence. In case the crew made an error or took alternative actions outside the scope of experimental procedures, an original EOP was available to the SRO for use as a contingency plan. However, the contingency plan was never required during the actual experiments.

2.5 Quantifying Sensitivity

Effect sizes were used to quantify the sensitivity of metrics. Cohen's *d* was calculated to compare the sensitivity of the WL metrics in different task types. We took the responses in checking tasks as a reference and compared the means from response implementation and detection tasks to the means from checking task. According to Cohen (1988), $d = 0.2$ is considered as a small effect size, 0.5 suggests a medium effect size, and 0.8 indicates a large effect size. Differences in WL across task types, especially the elevated WL of detection, have been robust across multiple studies. Thus, Cohen's *d* for task type comparisons would differentiate more and less sensitive measures, and the consistency of sensitivity differences across studies. Moreover, Analysis of Variance (ANOVAs)[3] were performed to compare the changes in psychophysiological responses and partial eta-squared was computed to illustrate the magnitudes of the effect sizes of task type manipulations measured by different psychophysiological metrics.

The sensitivity of WL metrics from the NASA-TLX, MRQ, ECG, EEG, fNIRS, and TCD was analyzed for Experiments 1-3, in which the subjective measures were administered after each task type block. In Experiment 4, only the psychophysiological metrics were available for task type-based comparison; therefore, the analyses for subjective measures (e.g., the NASA-TLX) were performed based on averaged means from three task type blocks in Studies 1-3. Due to lack of manipulation of task types, the analyses for subjective measures were less rigorous for sensitivity.

[3] Analysis of variance is a collection of statistical models and their associated estimation procedures used to analyze the differences among group means in a sample

3 RESULTS

3.1 Using Effect Sizes to Quantify Sensitivity

The sensitivity of metrics can be quantified using effect sizes. We used Cohen's *d* to compare the differences between means in task type conditions from available WL metrics. The responses in checking tasks were used as a reference to compare with the responses in implementation and detection tasks.

In addition to Cohen's *d*, ANOVAs were performed to compare the changes in physiological responses. The effect size measure used for the ANOVAs was partial eta-squared (η_p^2), which measures the proportion of variance explained by a given variable relative to the total variance remaining after accounting for variance explained by other variables in the model. The η_p^2 was computed to illustrate the magnitudes of the effects of task type manipulations measured by different physiological metrics.

3.1.1 Physiological Metrics

Table 5 summarizes the effect sizes for WL differences between detection or response implementation tasks and checking tasks. Generally, the effect sizes were larger in the comparisons between detection tasks to checking tasks than in the comparisons between response implementation tasks to checking tasks.

Oxygen saturation (especially in left hemisphere) measured by fNIRS showed consistent small to moderate effect sizes across studies 1-4, with the only exception for response implementation to checking comparison in study 4 in which the effect size was less than a small effect. In addition, EEG metrics showed consistent effect sizes across studies 2-4. Similarly, EEG responses in Studies 2 and 3 suggest that, with a touchscreen interface, the metrics are highly sensitive to the demands of detection. The analog, full-scope/scale simulator may have some elements of manual or physical involvement such as moving around to control the panels. EEG may pick up the same physical engagement which we observed from the touchscreen interface in studies 2 and 3 as well, but not necessarily with the desktop interface (study 1). EEG beta band and oxygen saturation in right hemisphere, it should be noted, seemed to be contradictory. Oxygen saturation measured by fNIRS may indicate overall WL, consistent with the higher WL which can be attributed to the vigilance tasks. EEG beta band may pick up some other task-induced response, such as mindlessness. People usually find it hard to stay focused as they daydream and mind wander. The EEG beta band may be picking up drift in focus.

Table 5. Physiological Measures Summary.

		Cohen's d (C – D)	Cohen's d (C – RI)
Study 1			
	HR	.422	.016
	HRV	-.235	-.187
	IBI	-.384	-.101
	Alpha	-.278	-.177
	Beta	-.020	-.061
	Theta	-.101	-.117
	SO_2-L	-.287	.228
	SO_2-R	-.534	.095
	CBFV-L	.268	.224

	CBFV-R	-.003	-.210
Study 2			
	HR	-.218	.262
	HRV	.505	-.479
	IBI	.237	-.275
	Alpha	.215	-.334
	Beta	.968	-.352
	Theta	.483	-.226
	SO_2-L	-.435	.271
	SO_2-R	-.423	.103
	CBFV-L	.081	.086
	CBFV-R	.194	.297
Study 3			
	HR	-.120	.385
	HRV	.051	.052
	IBI	.125	-.359
	Alpha	.355	.158
	Beta	.589	.211
	Theta	.362	.238
	SO_2-L	-.720	.121
	SO_2-R	-1.112	-.025
	CBFV-L	.039	-.224
	CBFV-R	-.556	.179
Study 4			
	HR	.410	-.387
	HRV	-.255	-.252
	IBI	-.521	.336
	Alpha	.536	-.394
	Beta	.682	.108
	Theta	.449	.774
	SO_2-L	.085	.165
	SO_2-R	-.398	-.155
	CBFV-L	.303	.532
	CBFV-R	-.259	-.515

Table 6 summarizes the means, standard deviations (enclosed in brackets) and effect sizes (η_p^2) measured by different physiological metrics across all four studies. This report focuses on sensitivity of the WL measures; detailed statistical analysis focused on overall performance and task-related factors can be found in previous reports (Reinerman-Jones & Mercado, 2014; Reinerman-Jones, Teo & Harris, 2016; Reinerman-Jones et al., 2018, 2019). It is clear from the data in the summary table that some measures were more sensitive than others, producing numerically larger effect sizes. For example, EEG beta showed consistent significant large effect sizes in studies 2 and 3. In both studies, EEG beta showed significantly smaller increase during detection tasks than other two task types. In Study 4 using analog, full-scope/scale simulator, a similar trend was revealed, suggesting that the participants may experience same effects in the analog, full-scope/scale simulator and the simulator with touchscreen interface, but not the simulator with desktop interface. Consistent with the Cohen's *d* results, oxygen saturation measured by fNIRS illustrated a similar pattern across the four studies. All the significant large effect sizes came from the highest ratings in the detection tasks. Although the results in study 4 were not significant, the detection task remained the one with highest WL ratings. This trend is consistent with the results of the detection task being the most demanding task as concluded in the previous reports.

Table 6. Physiological metrics summary. Asterisks denotes significant results.

	Checking				Detection				Response				Effect Size (η_p^2)			
	Exp1	Exp2	Exp3	Exp4	Exp1	Exp2	Exp3	Exp4	Exp1	Exp2	Exp3	Exp4	Exp1	Exp2	Exp3	Exp4
HR	5.41 (7.53)	2.51 (7.25)	6.84 (9.15)	11.66 (13.59)	3.06 (4.89)	3.45 (6.37)	7.70 (8.81)	8.38 (12.80)	4.83 (10.07)	1.29 (7.34)	4.29 (10.03)	12.57 (16.18)	.064*	.098**	.083	.188*
HRV	1.12 (22.84)	23.10 (35.41)	67.18 (151.77)	81.26 (125.03)	5.55 (18.95)	12.17 (30.38)	65.83 (149.96)	86.21 (119.35)	5.46 (30.11)	38.04 (39.38)	64.09 (151.10)	83.22 (133.01)	.025	.322**	.005	.008
IBI	-4.66 (6.82)	-1.97 (6.88)	-5.77 (7.90)	-8.81 (14.45)	-2.75 (4.76)	-2.98 (5.90)	-6.52 (8.23)	-6.11 (14.88)	-3.86 (8.14)	-.77 (7.11)	-3.28 (9.30)	-9.07 (15.83)	.067*	.112**	.087	.229**
Alpha	-21.56 (30.63)	3.73 (43.52)	7.27 (67.71)	81.67 (111.86)	-16.20 (24.95)	-1.31 (27.26)	-7.90 (41.95)	64.52 (101.00)	-13.50 (58.30)	21.27 (58.30)	1.61 (59.51)	76.79 (114.76)	.019	.115**	.079	.188**
Beta	14.72 (30.78)	76.83 (72.20)	63.44 (63.36)	132.54 (95.80)	4.10 (25.97)	34.35 (54.38)	32.03 (46.44)	116.31 (92.19)	18.66 (46.91)	96.35 (90.14)	54.59 (43.96)	130.14 (109.71)	.074*	.434**	.212*	.121
Theta	3.76 (28.96)	18.41 (33.76)	51.50 (102.67)	112.30 (130.19)	1.56 (26.73)	6.57 (19.34)	19.65 (46.98)	72.66 (99.74)	20.42 (114.24)	44.73 (120.16)	42.51 (76.77)	87.62 (108.42)	.027	.078*	.120	.285**
SO_2-L	-1.42 (2.75)	1.59 (4.30)	3.04 (3.02)	1.17 (2.42)	-.85 (2.65)	3.35 (3.42)	3.94 (2.53)	1.43 (3.09)	-1.85 (3.04)	1.04 (4.01)	2.79 (3.40)	1.25 (3.03)	.098**	.210**	.205	.040
SO_2-R	-1.06 (2.86)	2.22 (3.73)	2.13 (3.00)	1.66 (2.87)	-.09 (2.50)	4.14 (3.81)	3.71 (3.37)	2.11 (2.61)	-1.52 (2.86)	1.75 (3.91)	2.16 (3.77)	1.93 (2.92)	.214**	.263**	.589**	.089
CBFV-L	1.74 (7.72)	2.69 (8.83)	1.66 (6.96)	5.66 (16.87)	-.41 (7.21)	1.89 (9.82)	6.27 (4.72)	4.86 (16.74)	-.36 (8.28)	2.01 (11.43)	3.60 (9.60)	.32 (15.12)	.112**	.004	.226	.194
CBFV-R	-.56 (8.88)	2.26 (8.13)	1.57 (6.43)	4.17 (6.51)	-.92 (6.78)	.20 (10.14)	3.35 (3.77)	3.17 (7.14)	-2.20 (9.45)	.17 (7.66)	-.65 (12.12)	3.91 (7.31)	.026	.049*	.114	.073

3.1.2 NASA-TLX

Table 7 summarizes the effect sizes for pairwise comparisons between detection or response implementation tasks and checking tasks from participants' responses on the NASA-TLX in Studies 1-3. Due to the experimental design of using the original realistic EOP in Study 4, rather than using blocked steps in the same task type, the NASA-TLX was not administered after the task type manipulations. Therefore, the analysis of Cohen's *d* by task type was not available for Study 4, although we performed an analysis of overall subjective WL levels across studies. In the task type analysis, the effect sizes were generally larger in the comparisons between detection tasks to checking tasks than in the comparisons between response implementation tasks to checking tasks. Analyses revealed small or (near) large effects for global WL and the frustration subscale of the NASA-TLX across Studies 1-3. In addition, across studies 2 and 3 using simulators with a touchscreen interface, the effect sizes for the physical demand subscale were classified as in the medium range for the comparisons between the response implementation and checking tasks. Comparable effect sizes were not revealed for the physical demand subscale in study 1, which used a similar simulator with a desktop interface. The differences in sensitivity across simulator interface types suggests that the NASA-TLX could be used to determine the dependence of task effects on the particular control interaction technique used by the operator, such as the touchscreen.

Table 7. NASA-TLX Summary.

		Cohen's d (C – D)	Cohen's d (C – RI)
Study 1			
	Global workload	-.224	.069
	Mental demand	-.128	.104
	Physical demand	-.109	-.120
	Temporal demand	.275	.130
	Effort	.047	.019
	Frustration	-.711	.103
	Performance	.094	.079
Study 2			
	Global workload	-.266	.020
	Mental demand	.165	.102
	Physical demand	-.386	-.482
	Temporal demand	.177	.097
	Effort	.049	-.041
	Frustration	-.801	.081
	Performance	-.065	.123
Study 3			
	Global workload	-.925	-.085
	Mental demand	-.421	-.077
	Physical demand	-1.214	-.622
	Temporal demand	-.012	.294
	Effort	-.475	.104
	Frustration	-.705	-.274
	Performance	-.544	.091

Figure 1 illustrates a comparison of the overall NASA-TLX ratings collected after the entire experimental scenario in Study 4 to the averaged ratings from three task types in the first three

studies. As expected, the comparisons between student novices and former operators using the same type of simulator (Study 2 and Study 3) demonstrated generally lower WL perceived among the former experts. However, the effect sizes from ANOVA were modest, being less than .10 typically. Small effect sizes generally indicate overlap between two distributions. The error bars in the figure representing 95% confidence interval also suggest overlapping distributions The overlaps between distributions is an indication that the HPTF methodology of equal but different, where college student as "stand ins" for operators is feasible since the students experienced comparable levels of WL as the operators. Comparing the WL ratings from former operators using the analog, full-scope/scale simulator in Study 4 to the ratings in Studies 1 and 2 (e.g. global workload, mental demand, temporal demand, effort, performance), the WL levels were similar to the levels reported by the student novices. It seems that a carefully controlled experimental design reproduced some similar levels of cognitive demand in the realistic scenario performed by former operators.

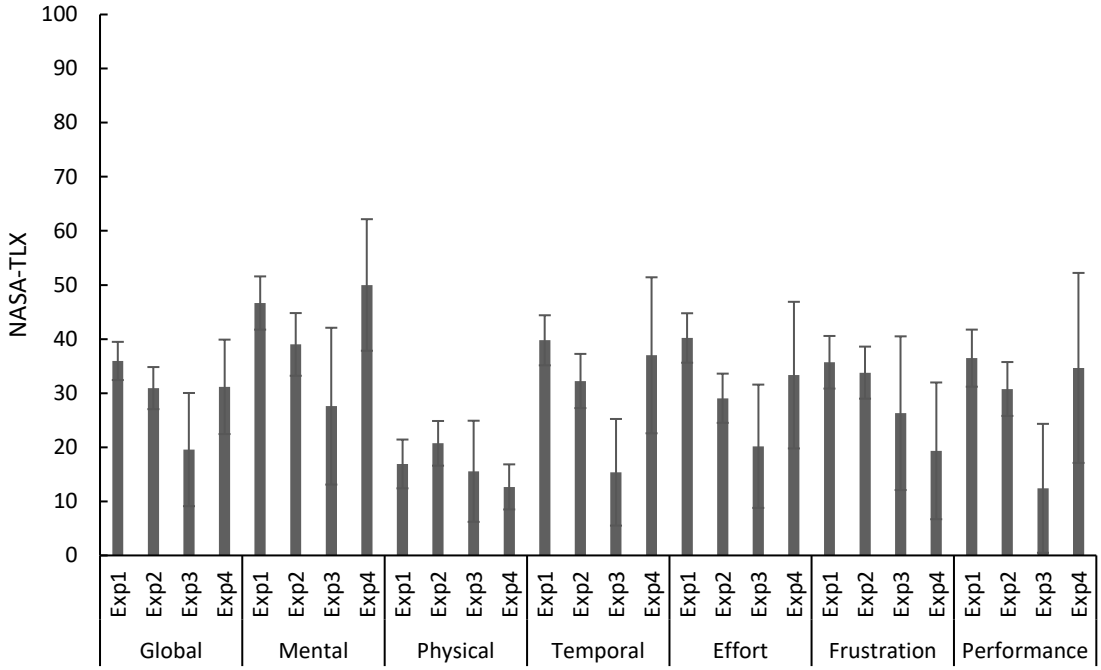

Figure 1. NASA-TLX ratings in studies 1-4. Error bars represent 95% confidence interval.

3.1.3 Multiple Resource Questionnaire (MRQ)

Table 8 summarizes the effect sizes between detection or response implementation tasks and checking tasks from participants' responses on the MRQ in Studies 1-3. Similar to the NASA-TLX results, MRQ data are not available for Study 4 because of the use of the realistic EOP, rather than task blocking. Therefore, the analysis of Cohen's *d* was not available for Study 4.

Several mental processes measured by MRQ were found to be sensitive to WL differences across the different task types. Analyses revealed meaningful effect sizes for spatial quantitative process (small to large effect), visual temporal process (near medium effect), and vocal process (small to medium effect) in the comparisons between detection tasks to checking tasks for Studies 1-3. By contrast, in Study 1, the effect sizes associated with the NASA-TLX data, indicated sensitivity to task-related WL differences for the frustration subscale and a small effect for global workload.

In addition, visual lexical process showed a consistent small effect when comparing pairwise between response implementation tasks to checking tasks in Studies 1-3. The comparison between response implementation and checking for studies using the touchscreen interface (Studies 2 and 3) and in the studies using college student sample (Studies 1 and 2) revealed that spatial categorical process exhibited consistent small to medium effects and manual process showed consistent small effects in Studies 2 and 3. Consistent small effects were also observed for spatial attentive in Studies 1 and 2.

Taken together, these data suggest that the MRQ, as evidenced by its sensitivity to task related variations in WL may perform better than the NASA-TLX from a diagnosticity perspective as it tends to identify specific sources of cognitive demand consistently across different interface types.

Table 8. MRQ Summary.

		Cohen's d (C – D)	Cohen's d (C – RI)
Study 1			
	Auditory Emotional	-.013	-.208
	Auditory Linguistic	.081	.005
	Manual	.143	-.077
	Short Term Memory	.100	.136
	Spatial Attentive	-.072	.242
	Spatial Concentrative	-.316	-.062
	Spatial Categorical	-.070	.061
	Spatial Emergent	-.065	-.041
	Spatial Positional	.113	.340
	Spatial Quantitative	-.348	.041
	Visual Lexical	.241	.192
	Visual Phonetic	.132	-.035
	Visual Temporal	-.415	-.093
	Vocal Process	.262	-.018
Study 2			
	Auditory Emotional	-.206	-.518
	Auditory Linguistic	.171	.249
	Manual	-.236	-.347
	Short Term Memory	.069	.055
	Spatial Attentive	-.034	.224
	Spatial Concentrative	-.069	-.079
	Spatial Categorical	-.361	-.356
	Spatial Emergent	.171	-.018
	Spatial Positional	.330	.092
	Spatial Quantitative	-.513	-.106
	Visual Lexical	.132	.244
	Visual Phonetic	-.040	-.356
	Visual Temporal	-.397	-.269
	Vocal Process	.435	.138
Study 3			
	Auditory Emotional	.003	.226
	Auditory Linguistic	-.030	-.002
	Manual	-1.173	-.486
	Short Term Memory	.069	.234
	Spatial Attentive	-.711	-.065
	Spatial Concentrative	-.710	-.150

Spatial Categorical	-.355	-.512
Spatial Emergent	.330	-.045
Spatial Positional	-.226	.270
Spatial Quantitative	-1.428	-.113
Visual Lexical	.186	.201
Visual Phonetic	.387	.128
Visual Temporal	-.589	-.237
Vocal Process	.639	.408

Figures 2 and 3 illustrate a comparison of the overall MRQ ratings collected after the entire experimental scenario in Study 4 to the averaged ratings from three task types in the first three studies. The trend in the MRQ subscales for spatial processes, such as spatial attentive process ($p < .01$, $\eta_p^2 = .087$), spatial concentrative process ($p > .05$, $\eta_p^2 = .019$), spatial categorical process ($p < .05$, $\eta_p^2 = .055$), spatial emergent process ($p > .05$, $\eta_p^2 = .016$), spatial positional process ($p < .01$, $\eta_p^2 = .082$), and spatial quantitative process ($p < .01$, $\eta_p^2 = .078$), were similar to the trend observed in the NASA-TLX. Former operators using the digital simulator with touchscreen interface reported the lowest ratings. A different trend was observed for MRQ ratings along the auditory dimension. Auditory emotional process ($p < .05$, $\eta_p^2 = .049$) and auditory linguistic process ($p < .05$, $\eta_p^2 = .049$), had higher ratings in student groups (Studies 1 and 2) than in former operator groups (Studies 3 and 4). The higher ratings may suggest that student novices had difficulty in dealing with communication in the context of the study tasks. In addition, the results from the visual temporal process subscale of the MRQ ($p < .05$, $\eta_p^2 = .064$) were consistent with the results from the temporal demand subscale of the NASA-TLX.

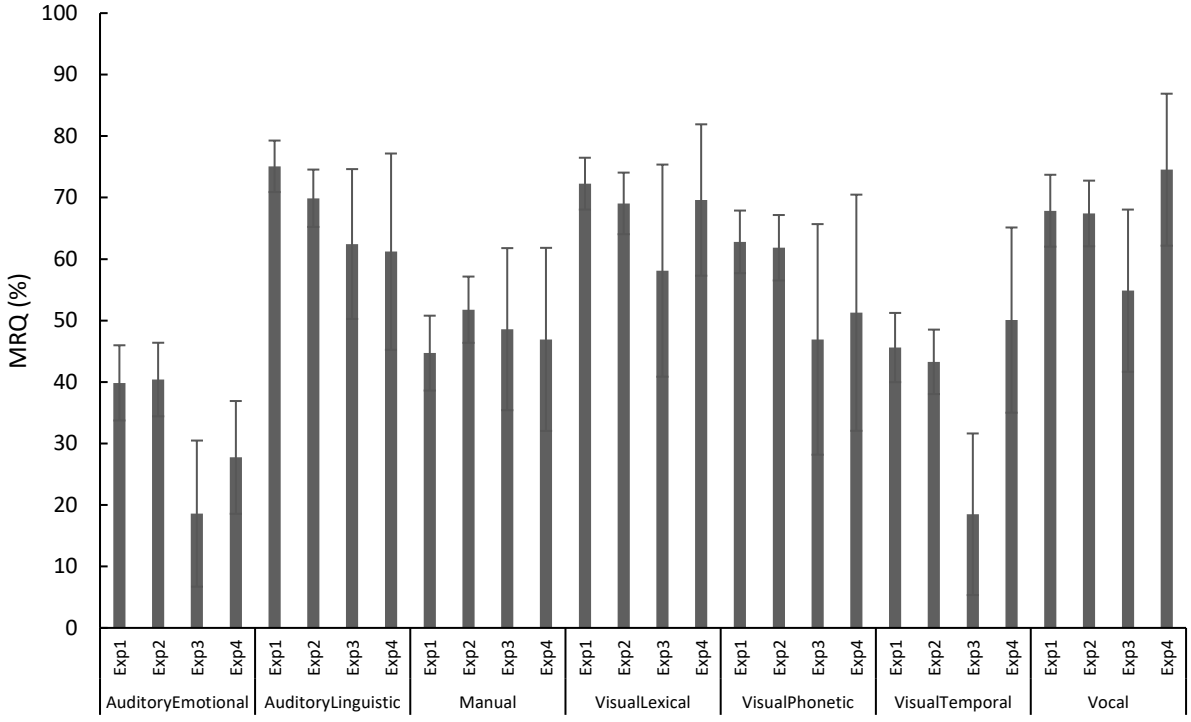

Figure 2. MRQ ratings in Studies 1-4. Error bars represent 95% confidence interval.

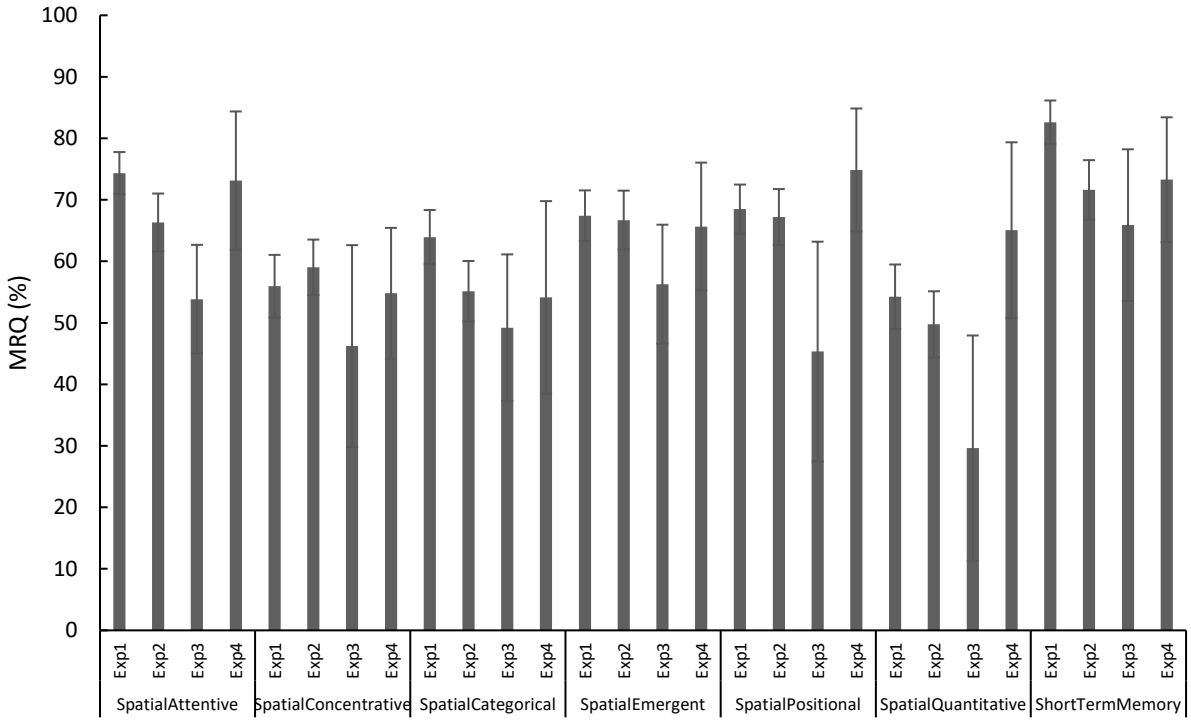

Figure 3. MRQ ratings in Studies 1-4. Error bars represent 95% confidence interval.

4 DISCUSSION

4.1 Sensitivity of Workload Measures

The analyses compared the effect sizes associated with the ratings captured by different metrics. Results indicate that workload (WL) responses induced by task type manipulations vary in magnitude, depending also on sample and interface type, suggesting sensitivity variation of the metrics in detecting WL changes depending on the assessment circumstances. The smaller effect sizes observed for some of the measures highlights the importance of robust *a priori* task definition and operationalization. RIL 2022-11 describes in detail the task analysis process used to define each of the common control room tasks used in the studies presented in the current report.

Psychophysiological WL measures are validated based on evidence of associations between variations in task demands and changes in physiological response. Psychophysiological measures are irreplaceable as they allow researchers to administer continuous measurement of participants' responses to experimental manipulations. In the HPTF studies, brain activity (electrical activity, cerebral flood flow velocity, and oxygen saturation) and cardiac activity were recorded and analyzed. Generally, the metrics of the psychophysiological WL measures are more sensitive in detecting the WL changes induced by the detection tasks compared to the response implementation tasks since the detection tasks are generally more demanding. In other words, from a resource theory point of view, detection tasks require more mental resource allocation, consistent with existing evidence (Warm, Parasuraman & Matthews, 2008). The similarly consistent EEG results from Studies 2-4 indicate that the touchscreen interface may

reproduce some comparable levels of human-system interaction which real operators experience in the MCRs with analog systems. EEG may pick up the similar physical engagement (e.g., moving around to manipulate I&Cs) involved in operations using analog and touchscreen simulators. ECG, as a means of unobtrusive and continuous cardiac monitoring, is one of the most common psychophysiological measurements of WL. Three metrics, heart rate, HRV, and IBI, derived from ECG were used in the studies. Unfortunately, no consistent trends were revealed to draw a solid conclusion in regard to ECG's sensitivity in profiling WL in nuclear domain. However, this does not suggest that ECG is never a good measure in practice. ECG is relatively easy to use and has been proved as a fairly reliable indicator of WL. Generally, an increase in mental WL leads to an increase in heart rate and a decrease in HRV (Veltman & Gaillard, 1996; Hankins & Wilson, 1998; Jorna, 1993). Divergence of cardiac data and other metrics is not uncommon. Literature shows that speech, respiration, muscle activity, body position, physical fitness, and many other factors can affect cardiac responses to WL manipulations in an experiment (Wilson, 1992; Jorna, 1992). In the HPTF studies, physical movement and frequent verbal communication were required (Study 1 participants in desktop condition were in seated position). Such physically diverse responses in a complex experimental scenario could mask small task-related variations in ECG. Thus, in practice when ECG or similar cardiac measures are used for any evaluation in the nuclear domain, assessment administrators must take all possible factors that may affect cardiac responses into consideration, if not able to control all the factors, and explain the results with caution.

Although psychophysiological measures may be more precise in some circumstances, subjective measures are more practical and easier to use. In addition, subjective measures reflect how participants feel about the experimental WL manipulations. In other words, subjective measures can reveal participants' perceived WL levels. Due to the limitation of the experimental design for Study 4, analyses for sensitivity of the subjective measures were only available among Studies 1-3. Similar to the general trend revealed by the psychophysiological measures, NASA-TLX and MRQ appear to be more sensitive in distinguishing the WL changes induced by detection tasks than the WL changes induced by response implementation tasks when comparing to the WL level induced by checking tasks. The MRQ may be more diagnostic than the NASA-TLX of specific sources of cognitive demand including the spatial quantitative, visual temporal, and vocal process demands common to both interfaces investigated here. The physical demand subscale of the NASA-TLX appeared to be reasonably sensitive to WL changes in both detection and response implementation tasks in the studies using simulators with a touchscreen interface. The difference in sensitivity in multiple studies using the same type of simulator suggests that the NASA-TLX could be useful in investigating human-system interaction involving different types of interfaces, such as touchscreens. More discussion regarding the NASA-TLX is covered in the next section. The MRQ as a measure to characterize the nature of the mental processes used during a task was found to be sensitive in profiling the WL differences induced by task type manipulations across multiple studies. Multiple subscales showed similar trend consistent with the trend revealed by the NASA-TLX. Such convergence of the subjective measures adds supportive evidence to the NASA-TLX as a reliable tool for practical use.

4.2 NASA-TLX

The NASA-TLX was initially developed for use in the aviation domain, but it has become widely used in other environments, such as nuclear power plants, military vehicles, unmanned systems, etc., as well. In the nuclear domain, NASA-TLX is used to assess individual's perceived WL levels in both controlled human-in-the-loop experiments in laboratories and for

operator evaluation during validation activities in nuclear power plant simulators. The NRC developed a generic metrics catalog and decision-making wizard to provide a guide to the NRC technical review staff to choose appropriate WL, situation awareness, or teamwork measurement for specific application reviews. Detailed information regarding this generic metrics catalog and decision-making wizard can be found in the NRC NUREG/CR-7190 (Reinerman-Jones, Guznov, Tyson, D'Agostino, & Hughes, 2015). As summarized in the NUREG/CR-7190, despite the advantages of high reliability, simplicity of administration, and non-intrusiveness (Farmer & Brownson, 2003), the NASA-TLX is subject to participant bias. It also does not provide a real time, continuous WL index. Moreover, NASA-TLX as a subjective scale often fails to show strong convergency with other psychophysiological and performance-based measures. Such lack of convergency of NASA-TLX and other metrics may result in confusion for technical staff about when to trust and when not to trust the results from NASA-TLX. Table 9 summarizes the results from selected psychophysiological metrics as well as the combined global WL and six subscales from NASA-TLX across Studies 1-3. Comparison of the subjective and objective WL measures shows evidence for both convergence and divergence.

Across all three studies, all significant WL changes detected by NASA-TLX (except temporal demand subscale in Study 1) converge with the oxygen saturation measured by fNIRS suggesting that detection tasks are more cognitively demanding than checking or response implementation. In Studies 1 and 2, effect sizes were substantially higher for fNIRS than for the NASA-TLX, suggesting that for these samples, the objective measure was more sensitive to inter-task differences in WL than the subjective measure. However, in Study 3, both fNIRS and NASA-TLX showed large effect sizes for the task type effect. Examination of the effect sizes suggests that the two measures are differentially sensitive to sample effects. For the two touchscreen studies, the fNIRS response is similar in both the novice (Study 2) and former operator (Study 3) samples, specifically, both groups produced elevated SO_2 during the detection task, relative to the other tasks. By contrast, global NASA-TLX WL for checking and response implementation was substantially lower in Study 3 than in Study 2, whereas WL for detection was more modestly reduced in Study 3 compared to Study 2. The especially low levels of WL experienced by former operators during checking and response implementation drive up the effect sizes in Study 3 by increasing the magnitude of the difference between the group means for those tasks. Thus, the NASA-TLX seems to be more sensitive than fNIRS to the benefits of experience. Alternatively, it might be argued that the cost of experience is loss of subjective awareness of cognitive demands, given that the fNIRS data suggest that neurocognitive demands remain across all three task types. From an evaluation standpoint, the data suggest that the NASA-TLX adequately captures WL differences between task types in experienced operators. However, the fNIRS may have superior sensitivity in novices.

The data in the table also suggests some degree of convergence between some NASA-TLX subscales and the trends in ECG results. In Study 1, NASA-TLX indicates that the checking task was perceived as the most temporally demanding, a finding which converged with data for heart rate and CBFV in the left hemisphere. The convergent data from multiple physiological metrics with NASA-TLX provides a strong indication that NASA-TLX is adequately sensitive in the specific context of novice participants performing with a desktop interface.

By contrast, Study 2 showed a distinctive WL response for response implementation, compared with the other two tasks, evident in higher HRV, and higher EEG power in alpha, beta and theta bands. There is no counterpart to these task type effects in the NASA-TLX data, implying that the psychophysiological metrics capture the impact of task-related WL on neurocognitive processing, which is likely not cognitively accessible and thus not reportable by participants on a

subjective WL questionnaire, like the NASA-TLX. However, although psychophysiological responses exhibit some level of convergence with NASA-TLX, these metrics can be affected by multiple factors as discussed earlier. Future research is needed to determine the extent of convergence and divergence between electrocardiac activity, EEG, CBFV and NASA-TLX, and the dependence of convergence on interface type and operator experience.

Table 9. NASA-TLX and physiological metrics convergency summary

	Checking			Detection			Response			Effect Size (η_p^2)		
	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3	Exp1	Exp2	Exp3
HR	5.41 (7.53)	2.51 (7.25)	6.84 (9.15)	3.06 (4.89)	3.45 (6.37)	7.70 (8.81)	4.83 (10.07)	1.29 (7.34)	4.29 (10.03)	.064*	.098**	.083
HRV	1.12 (22.84)	23.10 (35.41)	67.18 (151.77)	5.55 (18.95)	12.17 (30.38)	65.83 (149.96)	5.46 (30.11)	38.04 (39.38)	64.09 (151.10)	.025	.322**	.005
IBI	-4.66 (6.82)	-1.97 (6.88)	-5.77 (7.90)	-2.75 (4.76)	-2.98 (5.90)	-6.52 (8.23)	-3.86 (8.14)	-.77 (7.11)	-3.28 (9.30)	.067*	.112**	.087
Alpha	-21.56 (30.63)	3.73 (43.52)	7.27 (67.71)	-16.20 (24.95)	-1.31 (27.26)	-7.90 (41.95)	-13.50 (58.30)	21.27 (58.30)	1.61 (59.51)	.019	.115**	.079
Beta	14.72 (30.78)	76.83 (72.20)	63.44 (63.36)	4.10 (25.97)	34.35 (54.38)	32.03 (46.44)	18.66 (46.91)	96.35 (90.14)	54.59 (43.96)	.074*	.434**	.212*
Theta	3.76 (28.96)	18.41 (33.76)	51.50 (102.67)	1.56 (26.73)	6.57 (19.34)	19.65 (46.98)	20.42 (114.24)	44.73 (120.16)	42.51 (76.77)	.027	.078*	.120
SO_2-L	-1.42 (2.75)	1.59 (4.30)	3.04 (3.02)	-.85 (2.65)	3.35 (3.42)	3.94 (2.53)	-1.85 (3.04)	1.04 (4.01)	2.79 (3.40)	.098**	.210**	.205
SO_2-R	-1.06 (2.86)	2.22 (3.73)	2.13 (3.00)	-.09 (2.50)	4.14 (3.81)	3.71 (3.37)	-1.52 (2.86)	1.75 (3.91)	2.16 (3.77)	.214**	.263**	.589**
CBFV-L	1.74 (7.72)	2.69 (8.83)	1.66 (6.96)	-.41 (7.21)	1.89 (9.82)	6.27 (4.72)	-.36 (8.28)	2.01 (11.43)	3.60 (9.60)	.112**	.004	.226
CBFV-R	-.56 (8.88)	2.26 (8.13)	1.57 (6.43)	-.92 (6.78)	.20 (10.14)	3.35 (3.77)	-2.20 (9.45)	.17 (7.66)	-.65 (12.12)	.026	.049*	.114
Global workload	34.99 (16.97)	29.72 (16.79)	11.20 (8.83)	38.85 (18.90)	33.66 (19.58)	28.89 (21.43)	34.02 (19.53)	29.42 (18.63)	18.61 (15.27)	.048*	.048*	.452**
Mental demand	46.21 (25.20)	42.23 (25.99)	21.11 (19.17)	50.02 (28.32)	37.25 (28.46)	30.00 (21.94)	43.75 (28.88)	38.55 (29.23)	31.67 (27.95)	.025	.014	.114
Physical demand	12.84 (18.01)	15.80 (16.33)	4.44 (5.83)	20.59 (25.64)	23.26 (22.59)	33.33 (26.81)	17.31 (22.22)	23.12 (20.28)	8.89 (8.58)	.119**	.109**	.587**
Temporal demand	43.54 (23.95)	34.57 (24.69)	12.78 (15.83)	34.99 (28.99)	29.78 (26.90)	20.56 (17.04)	40.78 (25.56)	32.39 (25.62)	12.78 (17.70)	.049*	.017	.112
Effort	39.89 (24.41)	29.13 (21.78)	11.11 (9.93)	41.27 (28.99)	28.12 (23.47)	28.89 (22.75)	39.43 (24.99)	29.93 (22.30)	20.56 (15.70)	.002	.004	.467*
Frustration	29.14 (25.45)	26.38 (24.91)	12.78 (15.83)	51.26 (31.87)	50.29 (29.20)	40.00 (30.00)	26.73 (25.65)	24.71 (22.21)	26.11 (29.13)	.299**	.352**	.300
Performance	38.32 (30.26)	31.23 (27.77)	5.00 (6.12)	34.99 (29.45)	33.26 (27.29)	20.56 (29.42)	36.10 (31.64)	27.83 (25.65)	11.67 (16.01)	.005	.017	.238

5 CONCLUSION

The NRC's HFE Program Review Model, NUREG-0711, Rev. 3 (NRC, 2012) outlines that a generic "human centered" HFE design goal should include a design that supports personnel in maintaining vigilance over plant operations and provide acceptable WL levels. Furthermore, NUREG-0711's review elements highlight the importance of considering WL. For Elements 3 (Task Analysis) and 4 (Staffing and Qualifications) providing an estimate of WL is explicitly part of the review criteria that must be met.

Understanding the sensitivity of the WL measures used in the nuclear domain is important for numerous reasons:

- to enhance technical review staff knowledge of the utility and validity of these measures in a practical sense
- to enhance the technical bases of HFE program review guidance (NUREG-0711)
- to supplement the guidance documents where measures of WL apply
- to demonstrate the viability of using a multivariate assessment strategy for WL
- to supplement the empirical literature in the nuclear domain pertaining to WL measurement
- to further confirm the viability of the HPTF methodology and its ability to produce meaningful conclusions
- to better inform Human Reliability Analysis (HRA) model development.

Generally, the reanalysis of current data from the HPTF studies suggests that most of the WL measures utilized in the studies were sufficiently sensitive to the WL changes induced by the experimental manipulations in the simulated NPP operations that they would have practical utility. Some variation in sensitivity of the different measures was found, depending on sample and interface, but the higher WL associated with detection tasks relative to the other two task types was evident in multiple objective and subjective measures. These results indicate reasonable confidence that the measures used discern meaningful differences in terms of WL for control room tasks in a variety of contexts (e.g., novice vs former operator; interface technology, partial scale versus full scale simulator).

This report provides a summary of the convergence and divergence of the NASA-TLX with other WL measures which is important to NRC technical reviews because they most frequently encounter the NASA-TLX as a measure of WL proposed by applicants. In practice, the most widely used subjective measure is the NASA-TLX and the most commonly used physiological metric is heart rate. These measures exhibited some level of convergence, suggesting they are not only convenient to use, but also robust as reliable measures. The results of these analyses demonstrate good consistency among the measures such that technical review staff can have reasonable confidence in each of the measures analyzed. These results also indicate, however, that caution should be exercised in the interpretation of physiological results and the context of data collection should be considered. For example, given that cardiac activity is often affected by multiple factors, interpreting cardiac data needs to be done with caution, especially when verbal communication and physical movement is involved. A practical example where verbal communication and physical movement of the operators might come into play is in the Integrated System Validation which is demonstrated as part of the Verification and Validation element of NUREG-0711.

Convergence of NASA-TLX with oxygen saturation in prefrontal cortex is also a strong indicator of NASA-TLX being reliable in practice. Other findings suggested applications for WL measures beyond the NASA-TLX, with some of the physiological responses diverging from the NASA-TLX, depending on sample and task type. For example, EEG beta, contradicting oxygen saturation data, may reflect drift in focus. This may indicate an area where NASA-TLX is less diagnostic than some of the physiological measurements which may be able to detect some of these other more nuanced aspects of the complete “WL picture”. A multivariate assessment strategy is of particular importance in the nuclear domain because of the complexity of NPP operations and WL variation involved. Additionally, using multiple measures aids in hedging against the limitations of smaller sample sizes and the inherent complexity of assessing performance in an ecologically valid simulated control room environment. Convergence enables greater confidence in the trends in the data. Specifically, when multiple measures point to the same underlying WL component, confidence in the shape of the complete WL picture increases.

5.1 Implications for Human Factors Engineering (HFE) Guidance Development

Table 10 contains specific examples of NRC HFE guidance documents that may be enhanced or better informed by the results of the HPTF Volume 3 RIL 2022-11.

Table 10. NRC guidance documents where the results of these studies may be enhanced.

Guidance Document	Sections Related to Workload	Related Subsections
NUREG-0711 (Rev. 3) Elements	Task Analysis*	
	Staffing and Qualifications*	
	Human Factors Verification and Validation*	Integrated System Validation*
	Functional Requirements Analysis and Function Allocation**	
	Treatment of Important Human Actions**	
	HSI Design**	
	Procedure Development** (Described in SRP, Chapter 13 submittal)	
NUREG-1791 Staffing Requirements\Interim Staff Guidance Augmenting NUREG-1791, "Guidance for Assessing Exemption Requests from the Nuclear Power Plant Licensed Operator Staffing Requirements Specified in 10 CFR 50.54(m)," for Licensing Advanced Reactors under 10 CFR Part 53	Review the Task Analysis	
	Task Considerations	
	Implications for the review of exemption requests	
	Operational conditions sampling for advanced reactor design	
	Task Performance Requirements	
	Human Performance Measures and Criteria	
	Data Sources	Data from Human Performance Models
	Staffing Plan Validation Outcomes	

	Review Criteria for Operational Conditions	
NUREG/CR-7190		
	Supplement the empirical literature in the nuclear domain pertaining to workload measurement	
	Add references to these studies to the tool associated with NUREG/CR-7190	
	Generic Metrics Catalog (GMC) and Decision Making Wizard (DMW) (Microsoft Excel – 1.86 MB)	

* Workload Explicitly Mentioned

** Workload Not Explicitly Mentioned, however, it is Tangentially Related, and results could be considered

5.2 Implications for NRC Human Reliability Analysis (HRA)

Regarding enhancement of NRC HRA models, research on WL can enhance the formal process of HRA (Tran et al., 2007). Through a clearer understanding of the sensitivities of WL measures, these results help to determine how changes in task composition and interface design may impact vulnerability to human error. In particular, an enhanced understanding of the performance influencing factors of WL and task factors will help to better inform HRA methods. In addition to physiological WL measures as a potentially useful technique to enhance HRA data collection, the results of this research program can be used to inform dependencies between human actions. These same results could also inform the design and evaluation of the conduct of operations for current as well as new designs and concepts of operations (e.g., small modular reactors). For instance, future research can address how tasks should be allocated or distributed among crew members so as not to overload one individual. Modernization is likely to change teamwork both quantitatively and qualitatively. Fewer operators may be required as additional operator functions are automated, and what those operators do may also change. WL may become excessive if the capacity of automated systems to take over functions from humans is over-estimated, or if changes in plant design impair communication between operators or other teamwork activities. Designing and evaluating modernized plants requires attention to teaming issues, including teaming between humans and advanced automated systems.

6 REFERENCES

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.

Boles, D. B. (1991). Factor analysis and the cerebral hemispheres: Pilot study and parietal functions. *Neuropsychologia*, *29*(1), 59–91.

Boles, D. B. (1992). Factor analysis and the cerebral hemispheres: Temporal, occipital and frontal functions. *Neuropsychologia*, *30*(11), 963–988.

Boles, D. B. (1996). Factor analysis and the cerebral hemispheres: “Unlocalized” functions. *Neuropsychologia*, *34*(7), 723–736.

Boles, D. B. (2002). Lateralized spatial processes and their lexical implications. *Neuropsychologia*, *40*(12), 2125–2135.

Lin, J., Matthews, G., Barber, D., & Hughes, N. (2021). Comparing the Sensitivity of Workload Measures for Different Task Types Using Nuclear Power Plant Main Control Room Simulators. *AHFE Proceedings*. Applied Human Factors and Ergonomics Society Annual Meeting July 25-29, 2021, New York, NY.

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

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.

Boring, R. L. (2012). *Fifty years of THERP and human reliability analysis* (No. INL/CON-12-25623). Idaho Falls, ID: Idaho National Laboratory (INL).

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Farmer, E., & Brownson, A. (2003). Review of workload measurement, analysis and interpretation methods. *European Organization for the Safety of Air Navigation*, *33*, 1-33.

Hancock, P. A., & Matthews, G. (2019). Workload and performance: Associations, insensitivities, and dissociations. *Human Factors*, *61*, 374-392.

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.

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). Amsterdam, the Netherlands: North-Holland.

Henelius, A., Hirvonen, K., Holm, A., Korpela, J., & Muller, K. (2009). Mental workload classification using heart rate metrics. In *Engineering in Medicine and Biology Society, 2009: Proceedings of the Annual International Conference of the IEEE* (pp. 1836–1839). New York, NY: IEEE.

Hughes, N., D'Agostino, A., & Reinerman-Jones, L. (2017). The NRC's Human Performance Test Facility: Methodological considerations for developing a research program for systematic data collection using an NPP simulator. *Proceedings of the Enlarged Halden Programme Group (EHPG) meeting*, September 24-18, 2017, Lillehammer, Norway.

Jorna, P. G. A. M. (1992). Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological Psychology*, *34*(2), 237–257.

Jorna, P. G. A. M. (1993). Heart rate and workload variations in actual and simulated flight. *Ergonomics*, *36*(9), 1043–1054.

Leis, R., Reinerman-Jones, L., 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*, 210-214.

Matthews, G., Campbell, S. E., Falconer, S., Joyner, L. A., Huggins, J., Gilliland, K., ... & Warm, J. S. (2002). Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. *Emotion*, *2*, 315-340.

Mosleh, A., & Chang, Y. H. (2004). Model-based human reliability analysis: Prospects and requirements. *Reliability Engineering & System Safety*, *83*, 241-253.

O'Hara, J. M., & Higgins, J. C. (2010). *Human-system interfaces to automatic systems: Review guidance and technical bases* (BNL-91017-2010). *Human factors of advanced reactors (NRC JCN Y-6529).* Washington, DC: United States Nuclear Regulatory Commission.

Persensky, J., Szabo, Plott, C., Engh, T, Barnes, V. (2005). Guidance for assessing exemption requests from the nuclear power plant licensed operator staffing requirement specified in 10 CFR 50.54 (m) (NUREG-1791).

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). Heidelberg, Germany: Springer.

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.

Reinerman-Jones, L. E., Lin, J., Matthews, G., Barber, D., & Hughes, N. (2019). *Human performance test facility experiment 4: Former operator workload and performance*

comparison between two simulated environments. Rockville, MD: United States Nuclear Regulatory Commission.

Reinerman-Jones, L. E., Matthews, G., Harris, J., Barber, D., Hughes, N., & D'Agostino, A. (2018). *Human performance test facility experiment 3: Former operator workload and performance on three tasks in a simulated environment.* Rockville, MD: United States Nuclear Regulatory Commission.

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

Reinerman-Jones, L. E., Teo, G., & Harris, J. (2016). *Human performance test facility task order 1 technical report* (JCN # V621). Rockville, MD: United States Nuclear Regulatory Commission.

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

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.

Tran, T.Q., Boring, R.L., Joe, J.C., & Griffith, C.D. (2007). Extracting and converting quantitative data into human error probabilities. *Official Proceedings of the Joint 8th IEEE Conference on Human Factors and Power Plants and the 13th Annual Workshop on Human Performance/Root Cause/Trending/Operating Experience/Self Assessment*, pp.164-169.

U.S. Nuclear Regulatory Commission. (1983). *Handbook of human-reliability analysis with emphasis on nuclear power plant applications. Final report* (No. NUREG/CR--1278). Albuquerque, NM: Sandia National Labs.

U.S. Nuclear Regulatory Commission. (2006). Staff Requirements – Meeting with Advisory Committee on Reactor Safeguards (SRM-M061020). U.S. Nuclear Regulatory 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). 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?objectStor

eName=MainLibrary&ForceBrowserDownloadMgrPrompt=false&vsId=%7bA48838E6-D0A2-4969-B4CE-F168068567B6%7d

U.S. Nuclear Regulatory Commission. (2012). *Human Factors Engineering Program Review Model* (NUREG-0711, Rev.3). Washington, DC: United States Nuclear Regulatory Commission.

U.S. Nuclear Regulatory Commission. (2008). *Human factors considerations with respect to emerging technology in nuclear power plants* (NUREG/CR-6947). Washington, DC: United States Nuclear Regulatory Commission.

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. (2012). *Building a psychological foundation for human reliability analysis,* (NUREG-2114). U.S. Nuclear Regulatory Commission. https://www.nrc.gov/docs/ML1131/ML113180490.pdf

U.S. Nuclear Regulatory Commission. (2020). Integrated Human Event Analysis System for Event and Condition Assessment (IDHEAS-ECA). Research Information Letter (RIL 2020-02)

U.S. Nuclear Regulatory Commission. (2020). *NUREG-2198 The General Methodology of an Integrated Human Event Analysis System (IDHEAS-G)* (NUREG-2198). U.S. Nuclear Regulatory Commission. ML20329A428

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

Wilson, G. F. (1992). Applied use of cardiac and respiration measures: Practical considerations and precautions. *Biological Psychology*, *34*(2), 163–178.

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.