INL/EXT-14-31511
Rev. 1

Light Water Reactor Sustainability Program

Operator Performance Metrics for Control Room Modernization: A Practical Guide for Early Design Evaluation

March 2015

U.S. Department of Energy

Office of Nuclear Energy

INL/EXT-14-31511
Revision 0

Operator Performance Metrics for Control Room Modernization: A Practical Guide for Early Design Evaluation

Ronald Boring, Jeffrey Joe, Thomas Ulrich, Roger Lew

March 2015

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Office of Nuclear Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

(This page intentionally left blank)

ABSTRACT

As control rooms are modernized with new digital systems at nuclear power plants, it is necessary to evaluate the operator performance using these systems as part of a verification and validation process. There are no standard, predefined metrics available for assessing what is satisfactory operator interaction with new systems, especially during the early design stages of a new system. This report identifies the process and metrics for evaluating human system interfaces as part of control room modernization. The report includes background information on design and evaluation, a thorough discussion of human performance measures, and a practical example of how the process and metrics have been used as part of a turbine control system upgrade during the formative stages of design. The process and metrics are geared toward generalizability to other applications and serve as a template for utilities undertaking their own control room modernization activities. This revised report expands previous guidance on applying verification and validation approaches throughout the modernization design life cycle.

(This page intentionally left blank)

ACKNOWLEDGMENTS

This report was made possible through funding by the U.S. Department of Energy (DOE) Light Water Reactor Sustainability (LWRS) Program. We are grateful to Richard Reister of the DOE and Bruce Hallbert, Kenneth Thomas, and Kathryn McCarthy of Idaho National Laboratory (INL) for championing this effort. We also thank Kirk Fitzgerald, Brandon Rice, and Heather Medema of INL, who contributed their talents to making this project successful. Additional funding was made possible by the utility for which the technical work in this report was conducted. We are especially grateful to the program management, reactor operators, and simulator staff at the respective plants who contributed their time to this design phase of control room modernization.

(This page intentionally left blank)

CONTENTS

FIGURES

TABLES

ACRONYMS

AOP	Abnormal Operating Procedure
CFR	Code of Federal Regulations
DCS	distributed control system
DOE	Department of Energy
dpi	dots per inch
EPRI	Electric Power Research Institute
GVPC	governor valve position control
HFE	human factors engineering
HSI	human-system interface
HSSL	Human System Simulation Laboratory
I&C	instrumentation and controls
INL	Idaho National Laboratory
ISO	International Standards Organization
ISV	integrated system validation
LCD	liquid crystal display
LWRS	Light Water Reactor Sustainability
MCR	main control room
M&O	maintenance and operations
NASA	National Aeronautics and Space Administration
NRC	U.S. Nuclear Regulatory Commission
NPP	nuclear power plant
NUREG	Nuclear Regulatory Document
OP	operating procedure/output
OPC	overspeed protection control
PID	proportional, integrative, derivative/piping and instrumentation diagram
RO	reactor operator
RPM	revolutions per minute
SA	situation awareness
SART	Situation Awareness Rating Technique
SDS	steam dump system
SP	setpoint
SRO	senior reactor operator
TCS	turbine control system
TLX	Task Load Index
U.S.	United States
VCT	volume control tank
V&V	verification and validation

1. INTRODUCTION

Main control room (MCR) modernization is a reality at nuclear power plants (NPPs). With life extensions of plants beyond the original 40-year operating licenses, there is impetus to upgrade aging systems to achieve greater efficiencies and maintain high operational reliabilities. It is important that as these NPPs achieve license extensions, they continue to operate safely, reliably, and efficiently. Moreover, it is important that where they can gain efficiencies to maintain cost effectiveness, these efficiencies must be incorporated into the plants. Technology is one key source of efficiency, and it is common to see advances in electric production made possible through improved components like newer steam generators and steam turbines.

Upgrading control rooms is not an easy prospect. Several key challenges include:

- *The availability of spare parts for existing analog instrumentation and controls (I&C).* NPPs have out of necessity stockpiled multiple replacement parts for existing equipment in the control rooms. Broken parts are also serviced or rebuilt to extend their availability. These reserves are finite, but they have to date provided a steady supply to keep the plants functional, thereby obviating the immediate need for upgrades or new technology.

- *The availability of like-for-like replacement technologies.* While there are truly no remaining large-scale manufacturers of analog I&C, many vendors provide equivalent digital systems. For example, an analog gauge may be replaced with a digital system designed to accommodate the same inputs and provide equivalent output displays. These systems are essentially digital plug-and-play replacements for older analog technology, reverse engineered and designed to mimic the function and appearance of the legacy components as closely as possible. These technologies do not fundamentally change the control room but rather extend the life of the original design.

- *The limited offline time of the control room.* A typical NPP will operate around 18 months between refueling outages. During this 18-month period, many plants now operate the entire cycle without a single reactor trip. During refueling, the main control room is still the control center of the plant, along with an outage control center to coordinate maintenance and refueling activities across the plant. Because systems are constantly in use, there are very limited time windows in which to make changes to the control room. U.S. commercial plants in a deregulated energy environment would experience financial hardship to extend the outages in a manner that would allow significant change out of control systems in the control room. The large-scale control room modernization (with accompanying extended outages) witnessed in some European and Asian markets therefore does not readily translate to the U.S. marketplace. Control room modernization efforts must be accomplished quickly and on a small scale in the U.S.

- *The regulatory process of introducing new technologies.* A U.S. NPP is licensed to operate exactly in the manner it was built. Upgrades are changes to the plant, which typically require a license amendment. This process can be costly and time-consuming, and the approval of license amendments may not always be certain. This is especially the case in control room operations, which are integral to the safety of the plant and may garner extensive scrutiny before changes are allowed. Thus, it may be desirable for the utility simply to maintain the plant as-built rather than undertake a license amendment.

- *The training requirements for upgraded systems.* Licensed operators must be qualified through training to operate a new control system. This training is performed in the training simulator required at the plant. In order to facilitate such training, the new system must first be introduced into the

training simulator prior to implementation in the actual main control room. This sequencing must be performed in an expedient manner to ensure that all operating crews are adequately training without having a training simulator that is different from the actual main control room for any significant period of time.

- *The perceived limited return on investment for new control room technologies.* Most control room modernization has no effect on staffing levels in the control room. Unlike other upgrades at the plant, e.g., a turbine replacement system, there is no marked gain in efficiency, electricity generation, etc. In fact, control room modernization is a costly undertaking that promises minimal change in the overall operations of the plant. Reactor operators already operate the plant highly reliably and safely, and there would be no expected gains in reliability and safety. The key to achieving return on investment is twofold: (i) ensuring continued reliable operation of the plant through routine replacement of aging components including I&C in the control room and (ii) ensuring that new control systems improve on previous systems by aiding operators in monitoring, diagnosing, and controlling the plant. To wit, a hidden cost benefit of control room modernization is decrease in downtime for the plant. As aging control components require more maintenance, there is the potential for lost electricity production due to component failures in the control room. There have, for example, been cases when alarm system malfunctions have triggered temporary plant shutdowns while the alarms were repaired.[1]

- *The lack of experience in performing upgrades.* A hurdle to performing control room upgrades is the lack of industry experience in this arena. To date, few of the 99 NPPs in the U.S. have completed significant modernization of the I&C in the main control room. This lack of experience compounds the challenges above, because there is not always a clearly precedented path that the industry can take to move forward on control room modernization.

Since existing MCRs in United States (U.S.) plants are largely analog or mechanical systems and since equivalent analog or mechanical replacements for these systems cannot be readily obtained, modernization increasingly comes in the form of digital upgrades. In particular, utilities are replacing individual analog systems on the control boards with distributed control systems (DCSs) featuring digital displays, programmable logic control, and alphanumeric and touch input devices. These upgrades have to date been centered on non-safety systems, which do not require extensive license modifications through the U.S. Nuclear Regulatory Commission (NRC). Nonetheless, because the human-system interaction (HSI) between the operators and the DCS is considerably different than the analog systems it replaces, it is prudent to undertake a thorough process of ensuring the utility and performance of the new systems.

One of the key aspects influencing the effectiveness of the new DCS is the operator interaction with that system. Within the field of human factors engineering (HFE) is an area of specialization geared toward optimizing the design of the new HSI and assessing operator performance in using the new HSI. The U.S. Department of Energy (DOE) has established the Light Water Reactor Sustainability (LWRS) program to support research aimed as maintaining the current fleet of U.S. reactors through their life extensions. Among the areas of research within the LWRS Program is research centered on improving I&C, including the HSI. The Control Room Modernization Pilot Project works with utilities to conduct human factors research that helps utilities determine the best I&C upgrades to their control rooms. Since the MCR is

[1] The control room crew will err on the side of caution when there are component failures in the control room. In all cases, redundant systems ensure the plant can continue to operate safely, but it is industry best practice to repair faults rather than operate with workarounds. Similar practice is found in the commercial aviation industry, where planes are taken out of service for maintenance whenever a fault is detected.

heavily dependent on operator control, control room modernization especially benefits from the operator-centered emphasis of HFE.

Previous efforts under the LWRS Control Room Modernization project have developed a generic style guide for HSI upgrades (Ulrich et al., 2012); conducted the planning and analysis activities that are essential antecedents to new design work (Hugo et al., 2013); and developed a full-scale, full-scope, reconfigurable simulator capable of being used in control room modernization studies (Boring et al., 2012 and 2013). This latter effort is particularly noteworthy, as it provides a neutral testbed that may be used by utilities to support operator studies and basic design research necessary to transition to digital control rooms. The resulting Human-System Simulation Laboratory (HSSL) is depicted in Figure 1 in its current version. The HSSL currently supports four full plant models in a first-of-a-kind glasstop configuration that allows mimics of existing analog I&C as well as rapid development and testing of DCS technology on the virtual control panels. Individual collaborations with utilities are disseminated to ensure that HFE lessons learned benefit all interested parties, including other utilities considering control room modernization or the regulator that must review changes to control room functionality.

Figure 1. The Human-System Simulation Laboratory in Early 2014.

Because of the central role the operator plays in using the upgraded HSIs in the MCR, it is crucial that utilities properly design and evaluate their new systems using a vetted HFE process. However, currently available guidance on HFE for NPPs either does not extensively address control room modernization (instead focusing on new builds) or doesn't explain how to use an iterative design-evaluate process that provides early stage feedback on a novel design. This report is aimed at addressing gaps in the guidance for initiating control room modernization. Specifically, this report provides practical guidance and an example of the design and accompanying verification and validation of a new turbine control system (TCS) at an existing NPP. This report highlights a graded approach, in which only human performance measures that would readily be available at the plant are employed. This ensures the replicability of the process to other control room modernization efforts.

This report is divided into four chapters apart from this introduction:

- Chapter 2 provides necessary background on the design and evaluation process and methods
- Chapter 3 highlights the human performance measures that may be gathered as part of a design evaluation study

- Chapter 4 details an example of an early stage evaluation conducted on a new TCS design

- Chapter 5 summarizes the findings from the study and this report.

By providing practical guidance on early stage design evaluation in support of control room modernization, this report has three main objectives:

- To emphasize the importance of evaluation as an ongoing activity that supports design, not follows it

- To share a set of standard and easily used human performance measures that can support control room evaluations

- To demonstrate a graded approach to HFE in which a practicable, reasonable, and cost-effective process is used to support control room modernization.

In reviewing and referencing the insights from this report, it is believed utilities will find a systematic and readily extensible process that ensures the success of the HSI when embarking on control room upgrades.

2. DESIGN AND EVALUATION FOR CONTROL ROOM UPGRADES

2.1 NUREG-0711 Framework

The U.S. Nuclear Regulatory Commission (NRC) published the *Human Factors Engineering Program Review* Model in their NUREG-0711, Rev. 3 (O'Hara et al., 2012). The purpose of NUREG-0711 is to provide the procedure by which U.S. NRC staff review the effectiveness of human factors activities related to new construction and license amendments. Title 10, Parts 50 and 52, of the *Code of Federal Regulations* (10 CFR 50 and 52) provides the legal basis for requiring human factors considerations in nuclear power plant main control rooms. NUREG-0711 further defines human factors engineering as "The application of knowledge about human capabilities and limitations to designing the plant, its systems, and equipment." Put succinctly, NUREG-0711 outlines the process the regulator must follow to ensure that control rooms support the activities operators need to perform.

NUREG-0711, Rev. 3, contains four general categories of activities, ranging from planning and analysis, to design, verification and validation (V&V), and implementation and operation. Each of these phases is described below:

- The *planning and analysis phase* gathers information on the system, functions, tasks, and operator actions, which help to define the requirements for the system being implemented.

- These requirements, in turn, drive the second category of activities, related to *design* of the new or modified system. The requirements are turned into a style guide and specification and are then translated into the actual HSI.

- After the system design is finalized, it must undergo *verification and validation* to ensure that the system works as designed. Importantly, from a human factors perspective, the system should also be usable by the target users of the system, which are reactor operators in the case of the MCR. V&V remains an area of confusion in the field of human factors, as the distinction between verification and validation is not always clear. Fuld (1995) suggests that verification entails confirming existing truths, while validation confirms performance. This can be understood simply to mean that verification involves checking the HSI to an existing human factors standard like NUREG-0700 (U.S. NRC, 2002), while validation requires checking the performance of the system and operators according to desired performance criteria.

- Finally, the system must be *implemented and operated*, which includes monitoring operator performance in the actual use of the system.

These four main categories of human factors activities are further subdivided into a total of 12 elements, as depicted in Table 1.

While NUREG-0711, Rev. 3, is an invaluable guide to the regulator as well as a roadmap for many human factors activities by the licensee, it falls short of addressing three critical areas:

1. *Types of Testing Specified:* Chapter 8 of NUREG-0711, Rev. 3, outlines the required process for HSI design. The current version briefly references performing evaluations in the design phase—prior to V&V—but doesn't give detailed guidance. Specifically, Section 8.4.6 suggests there are two types of tests and evaluations that are appropriate at the design phase:

Table 1. The Stages of NUREG-0711, Rev. 3.

Planning and Analysis	Design	Verification and Validation	Implementation and Operation
HFE Program Management			
Operating Experience Review	Human-System Interface Design		Design Implementation
Function Analysis & Allocation	Procedure Development	Human Factors Verification and Validation	Human Performance Monitoring
Task Analysis	Training Program Development		
Staffing & Qualification			
Treatment of Important Human Actions			

- *Trade-off evaluations*, in which different design alternatives are considered, and
- *Performance-based tests*, in which operator performance is assessed.

These two are not mutually exclusive, e.g., performance-based tests can be used as part of trade-off evaluations. NUREG-0711 does not specifically require tests and evaluations during the design phase, nor does it provide examples of how such approaches are useful in shaping the design of the HSI. NUREG-0711 does require evaluation as part of the V&V activities conducted *after* the design phase. In particular, it advocates integrated system validation (ISV), which is "an evaluation, using performance based tests, to determine whether in integrated system's design (i.e., hardware, software, and personnel elements) meets performance requirements and supports the plant's safe operation" (O'Hara et al., 2012, p. 73). ISV is further elaborated in the earlier NUREG/CR-6393, (O'Hara et al., 1995). Note that NUREG/CR-6393 specifically states in Section 4.1.3 that the general evaluation methods used for ISV should not be used during earlier design phase activities, since they have different underlying goals. The ISV approach in NUREG-0711 and NUREG/CR-6393 has garnered criticism in terms of the limits of how well one set of test results can generalize to every possible subsequent situation (Fuld, 2007), an argument that could be extrapolated to suggest more frequent tests earlier in the process may generalize better. Still, an emerging consensus seems to be that verification works very well at the tail-end of design, while validation needs to be conducted earlier and iteratively (see, for example, Hamblin et al., 2013).

2. *Non-Safety Systems:* NUREG-0711 provides extensive guidance in Section 8.4.4.2 on control room requirements, but these requirements refer to overall systems—especially safety systems—that need to be present in the control room at design time. However, there is no guidance on individual non-safety systems. While non-safety systems (e.g., turbine control) are not subject to the same level of regulator review as safety systems (e.g., reactor control), a standardized set of good practices across both applications is desirable. There is no guidance on how to scale the approach to non-safety systems, including differences in the level of rigor expected.

3. *Modernization:* Finally, it must be noted that NUREG-0711 is optimized for reviewing initial license submittals (e.g., new builds) or license amendments (e.g., changing the operating characteristics of a required safety system). NUREG-0711 fails to provide clear guidance on modernization—replacement of an existing non-safety system—except to say that it should reasonably conform to operator expectations to minimize the need for additional training

Because guidance is missing on how to apply human factors engineering for modernization efforts on the existing fleet, the goal of this report is to augment the guidance in NUREG-0711 specifically to address how to upgrade existing HSIs for non-safety systems as part of a NUREG-0711 compliant (albeit unrequired) process.

2.2 EPRI Guidance

The Electrical Power Research Institute (EPRI) has published useful guidance on development of a human factors engineering process in support of control room modernization. *Human Factors Guidance for Control Room and Digital Human-System Interface Design and Modification: Guidelines for Planning, Specification, Design, Licensing, Implementation, Training, Operation, and Maintenance, TR-1010042* (EPRI, 2005) provides thorough discussions on a number of relevant steps in modernization, including control room modernization related to hybrid control room upgrades such as featured in the current LWRS projects.

Section 3.8 of EPRI-TR-1010042 emphasizes that these activities should be performed not as a single step after the design process but as a parallel activity coinciding with design. Important steps in the assessment prior to the final ISV include:

- Section 3.8.3.1: Planning for HFE V&V

- Section 3.8.3.2: Personnel Performing HFE V&V Activities and Criteria to be Used, verification activities performed by designers and validation by independent human factors experts

- Section 3.8.3.3: HSI Inventory and Characterization (e.g., location of displays, readability of graphical elements on displays, etc.)

- Section 3.8.3.4: HSI Task support Verification, in which representative tasks to be performed on the system are tested using operators using either static or dynamic HSI display elements

- Section 3.8.3.5: HFE Design Verification of the finalized HSI against design specifications and standards

- Section 3.8.3.6: Operational Conditions Sampling, in which key aspects of personnel tasks, plant conditions, and situations as determined in the planning and analysis phase (e.g., especially from the operating experience review) are tested

Within these suggestions, ERPI-TR-1010042 provides suggestions for performance measures in Section 3.10.3.6. These include measures to catalog the actions being carried out by the operators (e.g., responding to an alarm or navigating between displays), measures of task performance (e.g., time and accuracy to complete a given task of interest), and subjective measures (e.g., operator opinions on facets of the HSI).

EPRI-TR-1010042 provides helpful additional detail not covered in NUREG-0700, tailored to the specific task of control room modernization. It also emphasizes the importance of ongoing V&V activities as part of the design process, not simply as an end-state activity to be completed after the design is finalized and implemented.

2.3 A Simplified Framework for Testing Operator Performance

As noted, NUREG-0711 does not provide guidance for conducting HSI evaluations during the design phase. Here, we outline a simplified framework to redress this shortcoming and to provide the context and methods suitable for early stage HSI evaluation in support NPP control room modernizations. The key idea featured here is that of the iterative design cycle—one in which HSIs are designed, prototyped, tested, and improved in a cyclical fashion (Nielsen, 1993). Iterative design is central to the user-centered design process found in International Standards Organization (ISO) Standard 9241 that is at the heart of most human factors design activities (ISO, 2010). A core tenet of iterative design is that the resulting HSI will be more usable when built through an iterative process involving early testing rather than built to completion and then tested. Feedback provided early in the design process helps to ensure that error traps in the HSI are eliminated rather than ingrained in the design, meaning it is easier to fix usability issues earlier in the design than as a fix after the design is finalized. In terms of control room modernization, the equivalent argument would be that evaluation incorporated into the design phase will produce a system more acceptable, efficient, and useful to operators rather than one that features separate design and V&V phases. The approach we advocate includes a V&V activity at the end of the design process but also incorporates small-scale V&V activities in conjunction with design milestones. Thus, V&V becomes a staged activity rather than a single terminating activity after the design.

Figure 2 illustrates the idea of performing V&V activities prior to the formal ISV. In the depiction, the software specification and HSI style guide are developed based on information obtained in the planning and analysis phase. The software is then developed along three milestones during the design phase:

- At the first milestone (the 30% completion mark), the preliminary screen designs are completed. These screens can be evaluated as static (i.e. non-functional) screens, obtaining feedback from operators and experts on their impressions of the screen layout, look and feel, and completeness of information.

- At the second milestone (the 70% completion mark), the system dynamics are completed, and an initial functional prototype of the system may be evaluated by experts and operators. At this stage, operator performance may be assessed in use of the system.

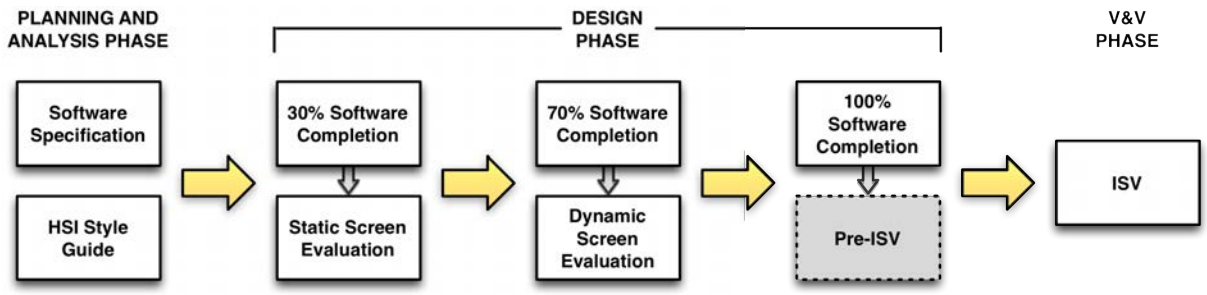

Figure 2. An Example of Design Phase Evaluations.

- At the final milestone (the 100% completion mark), the system may be tested a final time (in what might be called a dry-run or pre-ISV). Or, if there is sufficient confidence in the results of the two earlier evaluations, it may be appropriate to go directly to the ISV.

There are different verification vs. validation goals for the design phase and the formal V&V phase. It is useful to think of these two phases of evaluation as *formative* and *summative*. The notion of formative vs. summative evaluation is derived from the field of education (Scriven, 1967), where it is used to distinguish between methods to assess the effectiveness of particular teaching activities (formative evaluation) vs. the overall educational outcome (summative evaluation). The approach has been widely adopted in the human factors community (Redish et al., 2002):

- *Formative Evaluation*: Refers to evaluations done during the design process with the goal of shaping and improving the design as it evolves.

- *Summative Evaluation*: Refers to evaluations done after the design process is complete with the goal of confirming the usability of the overall design.

ISV is, by definition, summative, and it can be concluded that the guidance in NUREG-0711 is primarily of value to summative evaluations. What, then, of formative evaluations? Table 2 outlines different V&V methods suitable for formative and summative evaluation. Verification is accomplished by expert review against a standard set of criteria, while validation is performed via user testing. The following considerations apply:

Table 2. Usability 2x2 Matrix: Verification and Validation for Formative and Summative Evaluations.

		Evaluation Phase	
		Formative	**Summative**
Evaluation Type	**Expert Review (Verification)**	Heuristic Evaluation	Design Verification
	User Testing (Validation)	Usability Testing	Integrated System Validation

- *Formative Verification:* Completed during the design phase by expert review. Typical for this type of evaluation would be heuristic evaluation, which is an evaluation of the system against a pre-defined, simplified set of characteristics such as a usability checklist (Ulrich et al., 2013).

- *Summative Verification:* Completed after the design phase by expert review. Typical for this type of evaluation would be a review against applicable standards like NUREG-0700 (O'Hara et al., 2002) or requirements like the HSI style guide.

- *Formative Validation:* Completed during the design phase by user testing. Typical for this type of evaluation would be usability testing of a prototype HSI (ISO, 2010).

- *Summative Validation:* Completed after the design phase by user testing. Typical for this type of evaluation would be integrated system validation as described in NUREG-0711 (O'Hara et al., 2012).

In subsequent sections of this report, we focus on illustrating formative verification and validation to demonstrate both the process and utility of design stage evaluation of an HSI to be used in control room modernization.

2.4 Baseline vs. Benchmark

It is important to make a distinction between a performance *baseline* vs. a performance *benchmark*. The terms are often paired but used in vastly different domains. For example, a human resources definition would suggest that baselining is to compare current performance to historic performance, while benchmarking is to compare performance to others' performance (e.g., compare pay in one company to pay across the industry). More generally, while benchmarking implies a comparison (e.g., Boring et al., 2010), baselining does not necessarily require a comparison of different data points. Baselining can be an assessment of performance for a system at a particular point in time. The baseline measures can be used for trending, but they may also be used as standalone data. For the purposes of control room modernization, we define the two terms thus:

- A *baseline* is an evaluation of operator or system performance at a given point in time. A baseline may be used to evaluate the usability and ergonomics of an as-built system such as a particular HSI in the control room. Baseline findings may be used to catalog performance for use in longitudinal trending (over time) or to gather insights to inform the design of a replacement system.

- A *benchmark* is a comparative evaluation of operator or system performance. A benchmark may be used to evaluate the usability and ergonomics of two systems, such as when comparing an existing system vs. an upgraded system. Baseline findings may be used as part of a benchmark. A benchmark is often part of the validation of completed systems and is used to gauge the efficacy of a replacement system against its predecessor. In some cases, a benchmark may also be used to decide between competing prospective off-the-shelf system solutions.

For control room modernization purposes, the key distinction between a baseline and a benchmark is the stage at which it is employed. A baseline evaluation will be performed on an existing system *before* it is upgraded in order to inform the design of its replacement system. A baseline evaluation may also be performed periodically *after* a system is employed as part of maintenance and operations (M&O) to trend and ensure continued successful performance. In contrast, a benchmark is performed *during* the Design and V&V phases to ensure a new system performs at least as well as the system it is replacing. In human

factors terms, the benchmark ensures that the operators using the new system perform at least as reliably, efficiently, or safely as they did when using the predecessor system that is being replaced.

2.5 Gaps in NUREG-0711

2.5.1 Design Phase

In previous reports, we have discussed human factors specific to the Design phase of control room modernization. For example, an earlier version of this report (Boring et al., 2014) highlights operator performance measures that can be employed as part of design phase evaluations. We identified that the strict delineation between the Design phases and V&V phases overlooked an important opportunity for iterative design and evaluation. In other words, a good practice for human factors is not to complete the design and only then evaluate it. Rather, early design concepts should be evaluated and then refined, evaluated again, and the process repeated until a design with minimal operator performance issues is finalized. Throughout this process, prototypes should be used to afford rapid refinement and redesign as needed (Boring, Joe, and Ulrich, 2014; Lew et al., 2014; Ulrich et al., 2014). Only after the design is finalized and implemented is the formal ISV prescribed in NUREG-0711. There is a clear delineation between the Design phase and the formal V&V phase, but V&V is indeed necessary and desirable at the Design phase to help arrive at the final design. As described in earlier in this report, there is a need for formative evaluation of the interface during the Design phase, coupled with summative evaluation of the completed design prior to implementation. Formative evaluation is used to help shape the design, while summative evaluation is used to validate the finalized design. NUREG-0711 only addresses summative evaluation at length, but the utility will greatly benefit from using formative evaluation throughout the design cycle. As noted earlier in this chapter, human factors is considered least effective at the summative stage, when issues may prove entrenched in the design of the system and prove costly and time consuming to correct. Formative evaluation allows earlier discovery and correction of issues prior to implementation of the system.

The summative V&V phase in NUREG-0711 serves to document that the end product of the design operates as desired. Logically, the U.S. NRC, in reviewing licensee submittals related to control room modernization, is most interested in the results of the summative V&V in the form of the ISV study. However, while not explicated in NUREG-0711, an iterative design-evaluation cycle should be performed formatively during the Design phase to arrive at a satisfactory final design suitable for ISV. There may be reluctance on behalf of the utility to document to the regulator the findings of formative evaluations, since these evaluations will not represent rarified designs and will feature many issues that are ultimately resolved en route to the completed design. A design in progress is not a perfect design, and it is expected that there will be significant issues. Still, the fact that a systematic HFE process was followed to optimize the design is significant. The shortcomings of early designs should not be hidden; rather, there is value in documenting the evolution of the design. Even though documenting the evolution of the design through design-evaluation cycles is not required per NUREG-0711, the fact that such a process was followed lends considerable credibility to the final design.

2.5.2 Planning and Analysis Phase

Where we have previously espoused augmenting NUREG-0711 requirements for the Design and V&V phases to incorporate formative V&V, here we also point out that the Planning and Analysis phase has process gaps that need to be redressed from a utility perspective. The subelements within the Planning and Analysis phase of NUREG-0711—namely Operating Experience Review, Function Analysis & Allocation, Task Analysis, Staffing & Qualification Review, and Treatment of Important Human

Actions—are certainly applicable to control room modernization, as they are to new builds. As documented in Hugo et al. (2013) and depicted in Table 3, several of these subelements directly gather information that is useful to the design of the system. The importance of these subelements is not diminished for control room modernization applications. However, what is missing from the NUREG-0711 guidance, which is particularly relevant to control room modernization, is collection of baseline data.

Table 3. The Relationship between Planning and Analysis Subtasks to Design and V&V Activities.

	Operating Experience Review →	**Function Analysis and Allocation** →	**Task Analysis** →	**Design Activities** →	**Verification and Validation**
Goals	What happened before? Identify where existing system could be improved and where similar systems have provided relevant insights.	What is system vs. operator controlled? Identify opportunities to improve performance by indentifying modifiable functions.	What can be changed? Define information and control needs for operators to perform new and existing functions.	What's the new design? Develop conceptual designs for the HSIs.	Does it work? Test the designs and make sure all required information and controls are there and work.

Baseline performance evaluation, as noted earlier, entails collecting observations on how the existing system is used. The assumption here is that in control room modernization, there is not a need to hypothesize and determine the types of tasks operators will perform, because they are already doing them. Similarly, the upgrade to the control room will in most cases not introduce significant new functionality to the plant; rather, it will introduce new technology to the control room that will aid the operators in monitoring, diagnosing, and operating the plant. In some cases, additional functionality may be added, e.g., new sensors as part of a turbine control system upgrade may allow new control automation such as automatic synchronization to grid. However, new functionality represents the evolution of the existing process control, not the introduction of completely new processes. We certainly do not wish this to be a limiting statement. It is impossible to anticipate what new functionality may in the future be added to NPPs. However, at the current time in the U.S., the authors are not aware of any plant upgrades that would introduce significant new processes to the plant. Control room modernization is centered on upgrading existing, typically analog I&C to new digital technology, with only minor increments in automation or functionality.

As such, the design goals are not large departures from the existing system. For example, the Function Analysis and Allocation subelement of NUREG-0711 typically produces a functional hierarchy that includes components, systems, processes, safety functions, and goals. Where the purpose of a control room upgrade is to replace the I&C associated with a particular system in the plant, there would generally be no substantial change to the underlying components, systems, processes, safety functions, or goals. Therefore, a change in the HMI does not require a substantial reworking of the Function Analysis and Allocation associated with the predecessor system, unless significant new functions including new automation are planned as part of the upgrade. Similarly, if the overarching tasks performed by the operators are not significantly changed by the upgrade, the Task Analysis need only focus on those

changes associated with operators retrieving plant status information or performing control actions on the plant systems.[2] These are not new tasks, just refinements of existing tasks.

An alternate first step of control room modernization is a baseline evaluation of the current system already in place and currently being used. The baseline evaluation takes the form of a review of the usability and ergonomics of the current system.

- *Usability* is this case refers to the ease and reliability with which the operators perform required tasks. In order to conduct a usability evaluation, relevant scenarios related to use of the system should be selected and run in the plant training simulator or similar high fidelity simulator like the Human Systems Simulation Laboratory (HSSL) at INL (Boring et al., 2012 and 2013). The objective of the walkthroughs is to identify any opportunities for improvement in the HSI for the tasks performed by the operators. For example, a walkthrough of an existing turbine control system might note the requirement to have three reactor operators at the panels, because synchronization to the grid requires two operators at the turbine controls. Debrief interviews with the operators would identify why there is a need to have extra operators for that task and could identify particular tasks that are particularly resource demanding. Such information may be the basis for reevaluation of the Task Analysis or Function Analysis and Allocation. Ultimately, the usability evaluation will tell the design team what aspects of current operations are satisfactory, what information the operators rely on to complete tasks, and what improvements might be sought through an upgrade. These baseline data can also serve as comparison data points later when the replacement system is benchmarked against its predecessor.

- *Ergonomics* is the study of operators' physical interaction with the system. In this case, a baseline ergonomics evaluation will account for cases where physical strain is observed in or mentioned by the operators. For example, an operator might express that the process of closing a valve takes considerable time, during which the operator is unable to perform other tasks in the control room and the illuminated button actually becomes uncomfortably hot to the touch. Additionally, measures of the existing control boards should be taken and assessed to ergonomic standards like NUREG-0700, *Human-System Interface Design Review Guidelines* (O'Hara et al., 2002). The goal of the ergonomics review is to identify which areas of the physical layout of the boards (relative to the system being upgraded) are not optimized for use. In particular, the introduction of digital displays and input devices like trackpads to replace physical indicators, switches, and dials offers opportunity to consolidate the control boards. The ergonomics review will highlight areas where the consolidation should result in improved placement of sources of operator interaction with the system.

2.5.3 Revised NUREG-0711 Process Model for Control Room Modernization

Table 4 presents a summary of proposed additions to the NUREG-0711 HFE process model as proposed for control room modernization in this report. These changes are specific to the types of activities the plant should undertake as part of the HFE process, and they do not presume that such steps would necessarily be a required under regulatory review. In the Planning and Analysis phase, subelements for Baseline Usability Evaluation and Baseline Ergonomic Assessment are included. For the Design phase, the control boards must be reconfigured to accommodate the new digital control system, a design task requiring careful ergonomic review. This is represented as a new box entitled New Control Panel Layout. Also in the Design phase, an HMI Style Guide is added, which serves to direct the design elements of the

[2] Grandfathered systems may not have originally have undergone a NUREG-0711 process review, in which case the utility should undertake the Planning and Analysis subelements in order to align plant design documentation with current standards.

replacement system (see Ulrich et al., 2012). Formative Evaluation is also added to account for the iterative design-evaluation cycle described in Section 2.5.1. For the V&V phase, a subelement called Summative Benchmark is added, in which the baseline measures are compared to performance on the completed design. The Summative Benchmark is an appropriate treatment of ISV as described in NUREG-0711. No new subelements are proposed for the Implementation and Operation phase, but it should be noted that Human Performance Monitoring would resemble the periodic longitudinal baseline evaluations for M&O described in Section 2.4 of this report.

Table 4. NUREG-0711 Process Model with Added Steps Appropriate to Control Room Modernization.

Planning and Analysis	Design	Verification and Validation	Implementation and Operation
HFE Program Management	New Control Panel Layout*	Human Factors Verification and Validation	Design Implementation
Operating Experience Review	Human-Machine Interface Style Guide*	Summative Benchmark Evaluation*	Human Performance Monitoring
Baseline Usability Evaluation*	Human-System Interface Design		
Baseline Ergonomic Assessment*	Formative Evaluation*		
Staffing & Qualification	Training Program Development		
Treatment of Important Human Actions			

*Proposed additional activities by utility in support of control room modernization.

These modifications are neither expected nor endorsed by the U.S. NRC. It is our belief, however, that these changes represent appropriate additions to the HFE process for the utility to perform as part of control room modification. The additions complement the process outlined in NUREG-0711 and strengthen the HFE process in two critical ways:

- These new subelements are relevant to the utility. Whereas NUREG-0711 is geared primarily for U.S. NRC use and focuses on summative documents, these additional tasks ensure completeness of the HFE process by the utilities through the formative stages of control room modernization.

- NUREG-0711, as noted, is largely geared toward new builds. These steps help customize the HFE approach to the requirements of control room modernization.

3. HUMAN PERFORMANCE MEASURES FOR CONTROL ROOM VALIDATION

3.1 Introduction

There is no publicly available and widely disseminated standard template of metrics available to utilities to help them determine what is satisfactory operator interaction with a new system being developed and implemented. Granted, regulatory guidance on HSI design (e.g., NUREG-0700) and HFE aspects of NPP design and operation (e.g., NUREG-0711) exist, and licensee applicants use these documents to guide their license submissions. However, these documents were meant to be used by the U.S. NRC to evaluate the HFE aspects of licensee applications, and as such, they do not detail the metrics that the licensee should use to assess usability issues. That is, it was not the U.S. NRC's intention to regulate the specifics of how utilities address their HFE challenges, but rather to provide a design standard and a framework/guidance to their staff on the HFE review process.

Yet, as utilities engage in control room modernization activities, they need to be able to evaluate operator performance as they interact with new and upgraded systems, especially if they are systems that fall within the purview of the U.S. NRC's regulatory review of the utilities' HFE program (i.e., the usability of new safety systems is assessed in both the HSI Design Review and the Human Factors V&V elements of the NUREG-0711 review process). Even if the system under development, or in the process of being upgraded, is not considered a system requiring license amendment or regulatory review such as is the case with non-safety secondary systems, it is nevertheless prudent for utilities to adopt a single set of metrics for use to evaluate whether all systems being upgraded have satisfactory operator interaction. This chapter presents a general overview of what metrics, measures, and criterion/standards a utility can select and use to assess usability issues in their control room modernization activities. This chapter will further show that the metrics to be selected and employed will depend greatly on a number of key factors that relate to regulatory considerations. These include the type of usability study being conducted, the usability study's goals, the identified users, the testing environment and equipment, and other resource constraints.

3.2 Common Measures of Usability

Although it might not be widely known in other fields, the study of technology's usability has been a major focus of investigation within the field of human factors. Over its nearly 75-year history as a field of research, human factors professionals have put forth a considerable amount of effort investigating how to design or make technology more "user-friendly" and have refined and improved the methods, techniques, and measures over this time. According to Tullis and Albert (2008; see also Dumas & Redish, 1999; Rubin 1994), there are now a number of commonly used measures of usability. They include:

1. **Task success (also called 'effectiveness').** As the name implies, this metric assesses whether the user (i.e., end-user of the technology and/or participant in the study such as a reactor operator) is able to accomplish their task(s) or goal(s). Task success can also be thought of as effectiveness, which is a key usability attribute of ISO 9241 (2010) in that designs that enable the end-user to successfully complete their task will be also deemed effective. Defining whether a task is accomplished or not is typically tied to the reason the technology was created in the first place. That is, a technology is typically invented to solve a problem, and the end-user uses that technology to solve that problem. As such, task success is often defined by whether the end-user was able to use it to solve their problem

successfully (i.e., effectively) or not. This implies that task success is a binary metric, but it should be noted that task success can also be assessed in terms of degree (i.e., level of success or effectiveness).

2. **Task time (also called 'time on task' and 'task completion time').** The amount of time it takes an end-user to complete their task is another useful measure that produces a number of important insights into the nature of the technology's or system's usability. Tasks completed more quickly directly relate to a key attribute of the usability construct in ISO 9241 (2010): efficiency. Moreover, if a technology does not or is perceived by the user not to expedite task accomplishment, it will adversely affect the user's experience with it, thereby affecting another key attribute of the ISO 9241 usability construct: user satisfaction.

3. **Efficiency.** As mentioned above, efficiency is one of the core metrics of usability in that it is explicitly called out in ISO 9241 (2010). How quickly *and* with what level of effort the end-user must exert to accomplish their task is a measure of their efficiency using the technology or system. Obviously, tasks that take a long time and require a lot of cognitive or physical effort are not efficient. It should also be noted that tasks that can be done quickly but still require considerable effort are inefficient; likewise, tasks that require little effort but still take a long time to complete can also be considered inefficient.

4. **Satisfaction.** The extent to which the end-user is pleased with their experience interacting with the system or technology is a measure of their satisfaction. Though this is a subjective measure, it is a key attribute of the ISO 9241 definition of usability in that it is important to consider not only the objective performance of the system or technology, but also the end-user's perceptions and expectations of the system's performance (e.g., actual task completion time and perceived task completion time). For example, there could be two systems or designs that can be shown objectively to be equally fast to compete the same task, but if the end-user's perception is that one system 'feels' faster than the other, they will report greater satisfaction with it and will be more likely to use it. Similarly, if the end-user has an expectation of how quickly a task should be completed and the technology is perceived to be consistently slower than this expectation, the end-user will be less satisfied and will likely rate the technology lower in its overall usability.

5. **Errors.** The end-user's performance against established criteria that define success and failure/errors is another common way to measure the usability of a system or technology. For purposes of this report, errors are both incorrect actions that contribute to task failure and failures to act when action would have helped avoid task failure. Measuring what errors end-users make helps the designer identify what aspects of the usability of their technology or system are causing confusion or creating uncertainty in the end-user with respect to what action he or she needs to take (including no action) to achieve task success.

6. **Learnability.** For any technology or system that is sufficiently new to an end-user, its learnability is an important metric to capture. That is, it is a hallmark of good design that a novice user of a technology is able to learn how to use the system expertly in a relatively short amount of time. A technology or system is considered to have poor usability if it takes the user a long time (i.e., longer than expected), and/or the learning process to become proficient is fraught with frustrations on the part of the end-user.

7. **Self-reported metrics.** Self-reported metrics include paper surveys and questionnaires, as well as end-user responses to interview questions and verbal comments made during focus group sessions. Though self-report metrics are not orthogonal to the other metrics in this list, they are nonetheless very helpful measures to collect, especially with respect to measuring the end-user's satisfaction, and other attitudes about the technology and its design. When using surveys and questionnaires, it is

important to consider which type of scale (e.g., Likert scales, Semantic Differential scales, etc.) to use, as well as at what points during the usability assessment to issue these metrics (e.g., pre and post task). Generally speaking, self-report metrics collect input on usability issues more directly from the perspective of the end-user than other metrics. Other metrics can sometimes measure what the human factors expert believes to be a usability issue, which may or may not be an issue from the perspective of the end-user. For example, a human factors expert may be concerned that the design of a technology causes the end-user to spend more time on a task than another design, but it may not matter to the end-user if that design forces them to spend a lot of time on the task if it is enjoyable, or if the extra time has no consequence with respect to the end-user accomplishing their goal. In such cases, the extra time on a task might not negatively affect the end-user's overall evaluation of the technology's usability.

8. **Behavioral metrics.** Tullis and Albert noted (2008; p. 167):

   > During a typical usability test, most participants do much more than complete tasks and fill out questionnaires. They may laugh, groan, shout, grimace, smile, fidget in their chair, look aimlessly around the room, or drum their fingers on the table. These are all behaviors that are potentially measurable and offer insights into the usability of the product being tested.

   It could be argued further that there are times where the end-user's behaviors offer different and potentially contradictory insights into the their evaluation of the technology's usability. This can arise in situations where there is pressure (e.g., cultural norms), or incentives for the end-user to provide feedback on the usability of the system in a socially desirable way, or in a manner that maximizes the reward the end-user expects to receive. As such, behavioral metrics are another important facet to usability assessment, but it requires the human factors expert to become a keen observer of both verbal and non-verbal human behavior, or to make audio-video recordings of the end-users during the evaluations (and then exert a considerable amount of time and effort to review the recordings in detail to extract the relevant data).

9. **Physiological metrics.** Physiological metrics are related to behavioral metrics in that they can capture what are sometimes deemed 'involuntary' behavioral responses to stimuli, but they require specialized equipment to measure these reactions reliably. Common physiological measures for assessing usability include electromyogram sensors, eye tracking, pupillary response, skin conductance, and heart rate. The specialized (and sometimes expensive) equipment needed to measure end-user's physiological response also cannot provide real-time feedback to the human factors researcher (given the current state of this technology), thus again requiring the researcher to spend a considerable amount of time and effort analyzing and calibrating these data after the experimental sessions.

10. **Simulator logs.** In many commercial human factors usability assessments, it is possible to collect what Tullis and Albert called "live website data." For the types of usability studies being conducted for the LWRS program, this translates into simulator logs, which capture the interplay and timing of the interactions between what the simulator is doing (e.g., during a pre-defined training scenario), and the operator's responses. This is a very important baseline measure to have when trying to align all of the other data captured using other metrics (e.g., the operator grimaced *x* mins, *xx* secs into the scenario when he or she was unable to actuate safety injection).

11. **Combined issues-based metrics.** For purposes of this report, issues-based metrics are in its simplest form the aggregated frequency count of all of the usability issues the other metrics detect. This is a useful way of getting a more holistic picture of the usability of the technology or system. Furthermore, this aggregate metric is the starting point for the human factors research being able to

compare in a relatively simple fashion the usability of two or more technologies (or two or more designs) that are being evaluated. Of course, how the human factors researcher combines the issues identified adds to the complexity of how these issues are compared. For example, instead of weighting all usability issues equally, some issues may warrant being weighted more heavily (e.g., errors) than others (e.g., satisfaction) due to their significance with respect to the usability study's goal (e.g., regulatory acceptance of design). Further guidance on combining usability metrics and ways to compare findings can be found in Tullis and Albert (2008) as well as other common usability sources, but clearly other factors, such as what dimension(s) of usability the human factors researcher wishes to contrast and regulatory considerations, also need to be considered.

3.3 Selecting Measures

Despite the plethora of usability metrics that are available for use, it is also important to keep in mind that not all of these measures are applicable to the kinds of usability studies that can and should be conducted for NPP control room modernization. As stated above, the key selection criteria for which metrics a utility should use is the degree to which the usability construct being measured helps the design of the new system meet regulatory requirements. It is also no coincidence that the Usability 2 X 2 Matrix (see Figure 2) presented in Chapter 2 of this report presents a useful framework to help with the selection of usability metrics that help a designed and implemented system meet regulatory requirements. In particular, since this matrix identifies the different kinds of usability studies a utility might perform, it also identifies which logical metrics and/or the criterion (i.e., standards) used in the new system will help demonstrate an acceptable level of usability to the U.S. NRC.

Other important characteristics of the Usability 2 X 2 Matrix are consistency with a streamlined and graded approach to selecting metrics. For example, only certain kinds of studies can be conducted at certain points in the design process. Formative usability studies are done in the early stages of design, including conceptual prototypes. Summative usability studies are done in the later stages of design. As such, depending on where or at what stage the utility is in their development, some usability studies and their companion metrics and/or criterion are appropriate while others are not. Furthermore, sometimes the utility's schedule requires the formative usability study to be conducted so early in the design phase that only certain metrics are applicable given the nascent state of the new system, and oftentimes the summative study is done so late in the end stages of design that only certain metrics are useful in terms of providing actionable feedback to the utility and designer. In these ways, this usability matrix not only helps select which metrics to use (i.e., helps streamline the selection process), but also helps determine the number and complexity of the metrics selected (i.e., a graded approach).

Using the Usability 2 X 2 Matrix as a means to downselect on measures to use results in the following general outcomes:

- Expert reviews, whether formative or summative, will primarily extract the expert's insights on operator performance. For example, the expert may determine that a particular aspect of the interface is more error prone, but this represents the expert's best estimate that an error would result, not empirical evidence derived from operators using the interface. Most of the usability metrics may in similar fashion be estimated by expert review.

- User testing may avail itself of all the usability metrics. However, for formative evaluation, which may not involve a fully functional prototype of the end interface, it is often desirable to focus on metrics that capture basic performance (e.g., task success, task time, errors, efficiency, and learn-ability), often employing simple observation by human factors experts as well as self-reports by the

users. Unless there is a particular reason to suspect additional insights would be afforded by behavioral or physiological measures, these would typically be omitted during a formative test. Additionally, because the fidelity of the prototype system may be limited at the formative stage, some measures like task time may not be realistic at this stage. Because the formative evaluation seeks to discover user reactions to the system, measures may be centered on learnability and self-report, two metrics that are particularly useful in shaping design refinements.

- A summative review may require a higher degree of veracity, more akin to the factory acceptance test completed for hardware components. As such, the analysis is more detailed and may establish formal acceptance criteria. For example, the summative evaluation may seek to verify that users are able to complete a particular task successfully within a given amount of time. During the summative evaluation, the emphasis shifts from discovering what the users think about the interface to measuring their performance using the completed or nearly completed system implementation. At the summative stage, the emphasis is on objective measures—as such, observable behavioral or physiological metrics may be used. Since the system is fully functional, it is also possible to use simulator logs and combined issues-based metrics.

- For summative expert reviews, the criteria also become more objective—the expert is not asked to assess if he or she believes users might have issues with the color palette, for example, but rather to assess formally if the implemented color palette is compliant with applicable standards (e.g., NUREG-0700) and HSI style guides.

Some usability studies, such as usability testing and ISV, require more resources than heuristic evaluation or design verification. Usability testing and ISV both require knowledgeable operators as participants in the evaluation. Usability testing and ISV require at least a part-task simulator, and produce better results as the fidelity and scope and size of the simulator increases. These resource constraints on what study can be run also helps streamline and provide a graded approach to the selection process of metrics. In general, the following additional considerations should be factored into the selection of usability metrics:

- The Goals of the study.

- Regulatory considerations with respect to which metrics will demonstrate satisfactory operator interaction with the new system.

- Availability and/or accessibility of technology and experimental equipment to conduct the usability study (e.g., it may not be feasible to bring some physiological measurement equipment such as eye tracking into a simulator).

- Budget, schedule, and availability of expert end-users (e.g., licensed NPP operators), and human factors experts to conduct the study.

3.4 The Value of Simple Measures

The use of the specific measures in V&V is sometimes driven by the state of the art in human factors, not by their practical utility. This statement should not be misinterpreted to be a criticism of the many solid human factors approaches represented in the literature and in successful everyday use. There is a need for better measures, whether to refine existing measures or develop new ones. But, the fundamental question remains: *Are we actually measuring what we need to in order to perform the V&V?*

At a superficial level, the purpose of V&V is to establish that operator performance while using a system meets a minimum standard. That minimum standard may be set in terms of safety, reliability, workload, or other measures. The challenge is that these standards—and how to measure them—are not always clear. We need to do more work to establish the expectations of acceptable performance so that V&V studies can benchmark to that level. Without such clear standards, we risk the distractions of measurement novelties. Situation awareness, eye tracking, and physiological measures—while certainly constructively pushing the bounds of psychological measurement—may prove to be surrogates for the measures we actually need for operator performance. Sometimes straightforward usability measures such as described in this report and in Tullis & Albert (2008) may serve the needs of HSI evaluation adequately.

Again, our intent is not to criticize research that uses these types of measures, which may in fact be the key to understanding operator performance better. However, any number of advances in psychological measurement do not necessarily help us perform V&V better than we currently do. We must stop, catch our breath for a moment, and determine how different measurement tools available to us as researchers and practitioners actually help us understand operator performance. If our measures do not specifically verify or validate, we should discard or refine them. We must not be distracted by a gluttony of measurement.

The term *physics envy* has been suggested to describe the desire of so-called softer sciences to achieve the mathematical precision and lawlike certainty of physics as a hard science (Mirowski, 1999). It is a variant of this physics envy that underlies the two measurement issues:

- First, V&V researchers strive for a sort of finality in the findings of V&V. There is a desired precision and conclusiveness in saying a system has been verified and validated. It suggests that there's no room for error or refinement. The book is figuratively closed once the V&V is performed, and there's no need for questions. Perhaps this is one root of the tendency for late-stage ISV over early-stage V&V. This is akin to physics envy in the sense of striving for lawlike precision in our findings. Rarely are the findings from V&V so conclusive, even when they must stand up to regulatory scrutiny. We must learn to accept some degree of uncertainty in our V&V efforts. Humans are remarkably resilient to consistency and classification. We must make our peace with the imprecision of V&V. A better approach is to show the trajectory of the findings. This is demonstrated through iterative evaluations early in the design—showing the refinement of the system design and the improvement of operator performance while using the system. It is the process of improving the design—not the immutability of the V&V findings—that determines the system is successful and usable by operators.

- Second, as noted, V&V researchers have tended to gather increasingly complex measures of performance. It might be argued that this is in pursuit of a more scientific and scrutable set of findings rather than the subjective measures we must often employ in our studies. Certainly, the pursuit of better measures should be applauded. But, these measures must not be applied simply to further the hope of greater scientific precision. A good measure is any measure that provides insights into operator performance. It is not simply the quality of the measure but rather the quality of clearly matching the measure to V&V objectives that will ultimately prevail the science of V&V.

As V&V researchers and practitioners, we sometimes envy fields that provide highly conclusive findings, and we compensate with an ever increasing arsenal of measurement methods. Instead, we should embrace the evolving nature of the findings afforded by early-stage evaluation using relevant measures to support our analysis. We should refine our processes and measures to best reflect operator performance and system interfaces.

3.5 Knowledge Transfer

Examining the design issues from the perspective of human factors can be a means to bring different areas of expertise together to discuss and resolve design issues. Human factors, by definition, frames engineering issues in the context of the human operator, who is an important component of the system (e.g., the last line of defense). Once the problem space is defined with the human operator as the common denominator, a number of important factors critical to effective and timely knowledge transfer including the development of a common language for all experts to use in describing the elements of the system fall into place. Timely and effective knowledge transfer can be conceptualized as the direct and well-coordinated feedback provided by the expertise of systems engineers, plant managers, and operators to the design team as part of the V&V process.

Knowledge transfer is a direct outcome of V&V; however, from the larger perspective of industry-wide control room upgrades, knowledge transfer refers to the well-coordinated and effective exchange of information between all the different entities involved so that future modernization efforts can be more effective. Timely and effective knowledge transfer between the different experts can help team members understand how their particular expertise can augment the overall design team's collective expertise.

Thus, from a larger design process perspective, the human factors experts can serve the crucial role of objectively identifying the strengths of the different specialties and facilitating the interactions between the experts of these different specialties. This benefit is sometimes overlooked, but overall, we believe that when V&V produces successful outcomes, the important content (e.g., design input) and process contributions (e.g., developing a common language and providing a "neutral" playing field) supersede the value of individual measures used.

(This page intentionally left blank)

4. HUMAN PERFORMANCE MEASURES EXAMPLE

4.1 Introduction

A three-day formative evaluation workshop was conducted at the HSSL for a turbine control system (TCS) interface under development for an existing fleet of nuclear reactors belonging to a U.S. based utility (see Figure 3).[3] The fleet reactors for which the TCS is being developed are a mix of pressurized water reactors and boiling water reactors with nominal 1000 MW power generating capacities. Each reactor has a multistage steam turbine with one high pressure stage and a low pressure stage with two turbines. Due to the proprietary nature of the TCS and plants, plant specific information has been omitted from this report. Despite this intentional sanitizing of information and results, the goal here is to provide the readers a descriptive account of the method and structure of the three-day workshop. Additionally, this section provides a high-level overview and review of the qualitative and quantitative measures collected during the workshop.

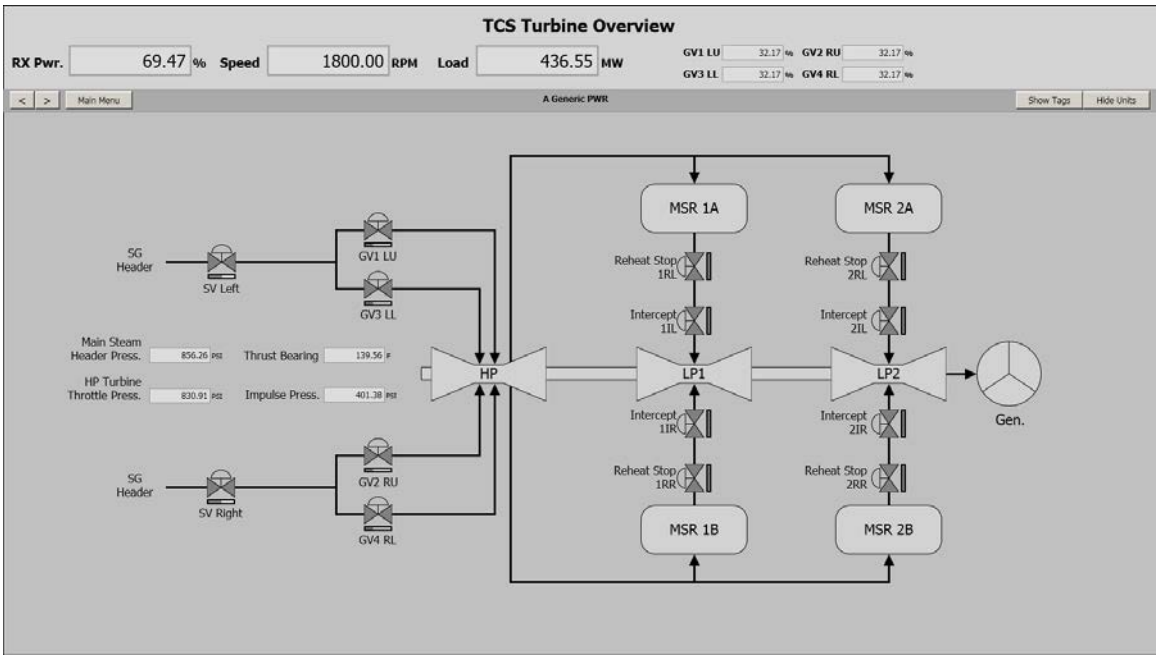

Figure 3. A Digital TCS Overview Screen with a Steam Flow Process Mimic.

In total 22 individuals from six entities participated in the workshop. These individuals brought a diverse set of domain knowledge, skills, and expertise that contributed to the overall success of the workshop. The utility provided three licensed power plant operators, various operational personnel, a plant instructor, and a project manager. During the workshop these human resources provided valuable knowledge regarding how their plant functioned, how it was actually operated on a daily basis, and the timeline for the project. The plant instructor's expertise running the power plant simulator was also extremely useful in setting up realistic scenarios to test the TCS. The TCS vendor provided a turbine

[3] To protect proprietary designs, the figures in this chapter were constructed for this report as a representative example of what a modern TCS interface might look like. They do not represent an actual TCS, either planned or currently in cooperation.

control system engineer who in turn provided nuanced distinctions between the existing TCS and the TCS under development. The actual TCS interface under evaluation was designed by yet another entity using Honeywell's Experion DCS platform. The two delegates from the interface design vendor contributed knowledge regarding the system's interface requirements and specifications, rationale for design decisions, as well as expertise regarding the Experion platform's capabilities and limitations. An independent consultant with 42 years of experience in nuclear human factors engineering provided unbiased expertise and valuable high-level insights. Lastly, the 11 individuals from INL and DOE contributed to setting up and carrying out the workshop. The full-scope fully-reconfigurable glasstop NPP main control room simulator at INL was used to simulate the native control room of our guest operators. In addition to providing this one-of-a-kind simulator platform, INL provided human factors expertise, nuclear instrumentation and control expertise, and the technical expertise required to support the nuclear plant simulator and accompanying control room.

The primary goal of the workshop was to evaluate and remedy shortcomings of the TCS interface under early stage development. User testing was used to identify whether critical information was missing and whether the interface contained misleading information or cognitive traps. For an interface to be successful, the representation it conveys needs to be compatible with the physical system as well as the operator's mental model. The information it presents should be pertinent, effective, and intuitive to the operator (Vicente, 1997). In order for the HSI designers to meet this criterion of success they need to understand the physical system and the associated mental models of operators for the system. In order for operators to effectively use the interface their mental model needs to be compatible with the physical system. In order for process engineers to design better processes they need to fully understand how operators control the plant. Pursuant to these needs a secondary goal of the workshop was to provide a forum for face-to-face discussion and collaboration—to bring the various mental models in synch and translate this into a coherent HSI. In pursuing those goals the workshop attained other benefits including the identification of potential logistical problems in upgrading the plant and the identification of potential problems in operating the plant with the new TCS.

4.2 Validation: Usability Testing

4.2.1 Method

A single three-member team of operators was exposed to two phases of the study during the workshop. In the first phase, the operators used the existing TCS interface, while during the second phase, the operators interacted with the TCS under development. This corresponded to the baseline and benchmark measures. Throughout these two phases of the study, the operators operated the simulated plant under different scenarios derived from training scenarios and designed to cover a representative spectrum of normal and off-normal turbine evolutions. The team was comprised of three licensed operators--one licensed senior reactor operator (SRO) and two licensed reactor operators (ROs). All operators had acquired more than 20 years of experience in the nuclear industry.

4.2.1.1 Procedure

On the first day of the workshop the operators conducted four scenarios using their current control room setup on the HSSL's full-scope glasstop simulator. The instructor from the plant directed the scenarios and instructed the operators to interact and behave as if they were conducting a routine training exercise. The plant simulator was running and provided the full plant dynamics of the various scenarios. These

scenarios served as baseline measures of the plant TCS as currently implemented. As previously mentioned, operators were intimately familiar with the simulated plant and control room layout. However, they had no previous experience using the touchscreen digital panel mimics. Here, it will suffice to say that the operators quickly adapted to the panels, and anecdotally the SRO remarked at the conclusion of the first scenario how surprised he was at how close it felt to the real plant.

On the second day, after being introduced to the new TCS (hardware, logic, functions) and the new TCS interface, the team walked through the same four scenarios with functional mockups of the new digital control system placed on revised panel mimics within the glasstop simulator. The mockup DCS screens were made navigable using INL's prototyping tool for rapid prototyping on the glasstop simulator (Lew et al., 2014). Operators were instructed to think-aloud as they ran through the scenarios. The operators' mental models of the plant, the TCS vendor's mental model of the new control system, the interface designers' expertise, as well as procedural notes from the previous day allowed the operators to visualize what they would need to check and control using the new interface and how the physical system would respond.

While the scenarios were being conducted two evaluators recorded time-stamped measures of operator actions and plant evolutions. A third evaluator operated a handheld camera while two additional evaluators and the plant instructor oversaw the technicalities pertaining to the simulator. On the first day, operators performed the scenarios without interruption followed by a detailed debrief. On the second day the nature of the scenario walkthroughs resulted in semi-structured discussions of the new TCS or TCS implementation logistics. The independent consultant present at the workshop acted as a facilitator to keep things moving along when he deemed necessary.

On the third day, the plant personnel walked through the 43 prototype screens to identify any issues with the screens. These issues included information that didn't align with their expectations, unconventional labels, and awkward navigation between screens. These issues resulted largely from the new TCS being adapted from a non-nuclear configuration.

4.2.1.2 Scenarios

Scenario 1: Steam Generator Tube Leak. The first scenario was initiated with the reactor online and the turbine at full load. A small but detectable steam generator tube leak was created in the plant simulator. The leak exceeds the volume control tank (VCT) makeup capacity and requires entering an abnormal operating procedure (AOP) to rapidly downpower the radioactivity. Reducing radioactivity requires operators synergistically reducing turbine load. Behind the scenes, the simulator operator gradually ramps up the leak until it reaches roughly the 40 minute mark, when it develops into a 400 GPM rupture, causing the operators to abort the rapid downpower AOP and trip the reactor. The scenario ends after the turbine operator verifies the turbine throttle and governor valves are closed.

Scenario 2: Synch to Grid. Scenario 2 required the operators to ramp up the turbine and synchronize the turbine to the grid. The scenario begins with the reactor at 7% power and the turbine offline with the turning gear engaged. With both the old and the new TCS, the operators must first complete a series of tests by checking various valve states before they can latch the turbine. Then they must perform an overspeed protection control (OPC) test before ramping up the turbine. With both the old and new systems they hold the turbine at intermediate hold points outside of critical turbine resonance ranges and verify critical parameters before increasing turbine speed until the turbine eventually reaches 1804 RPM. With the existing TCS, operators used the "Oper Auto" mode. They specified a demand speed in RPM and a demand ramp rate in RPM/minute and the TCS did the rest. With the new TCS operators will have an "Auto Start" mode that will hold at the correct hold points and will dynamically adjust the ramp rate

based on turbine resonance ranges and high pressure metal temperature. Rapid temperature changes of turbine parts can cause large internal temperature variations resulting in thermal stress and cracking. Once the turbine is at 1804 RPM, the generator is ready to synchronize to the grid. 1804 RPM ensures that the syncroscope is moving "slowly in the fast direction," slightly faster than the optimal temperature of 1800 RPM. When the syncroscope is close to the 12 o'clock position, closing the generator breaker will synch the generator to the grid. As the generator picks up load, operators must also monitor steam flow and primary power. On the first day with the plant dynamics running this scenario took 73 minutes to complete.

Scenario 3: Loss of Load. In the third scenario the plant was running at full power and at full load. In the plant simulator a ramping loss of load event was applied. When electrical load is dropped the turbine is prone to overspeed and trip in a matter of seconds if the main steam is not controlled by either raising the control rods thereby reducing reactor power and the produced steam, or dumping main steam to the condenser or atmosphere via the steam dump system (SDS). A load rejection controller monitors load and responds accordingly. With this particular plant, when the load rejection exceeds 5% of the load capacity, steam dump is actuated. The simulated loss was of sufficient magnitude to initiate a steam dump to condenser and produce a corresponding alarm. Operators entered the AOP for secondary load rejection. In the plant simulator, the load was restored and the load rejection controller closed the steam dumps. Once the operators realized the transient nature of the event they reset the steam dump system and the scenario ended.

Scenario 4: Multiple Faults. The final scenario put the operators' divided attention to the test by simultaneously presenting multiple failures. Two of the failures emulated a propagating turbine bearing vibration by gradually increasing the amount of vibration in the "#9" bearing and after a few minutes introducing a vibration in the "#8" bearing. A second set of failures emulated an oscillating governor valve indicator and a failed closed governor valve. The turbine vibration required entering the AOP for turbine eccentricity and vibration and operators diligently followed the flow paths while simultaneously trying to monitor the boards. Once all the failures were identified the scenario was ended.

4.2.2 Basic Collected Measures

We used a holistic data collection approach aimed at capturing as many aspects of the workshop as possible to maximize the research benefits from this workshop. As a result, some of the measures yielded immediate results, such as the expert review and discussion, which were put to immediate use by the HSI design team in the next iteration of the turbine control system. Other measures, such as the simulator logs, video recordings, and behavioral logs, are data rich and will require future extensive analysis before formal results are presented. A summary of the preliminary results for each measure and the current and future utility of those results are described in the following sections.

4.2.2.1 Behavioral Logs

During the four simulated turbine control scenario walkthroughs, several human factors experts recorded operator behaviors. The four scenario walkthroughs consisted of an exemplary plant team of operators comprised of a senior reactor operator and two reactor operators that physically manipulated the simulated control panel mimics required for turbine control as they are currently configured without a digital control system at the actual plant. To capture the behavior of the entire shift crew, one human factors expert focused on recording the procedure navigation, procedure actions, and verbal commands issued by the senior reactor operator. Two additional human factors experts focused individually on recording the actions and verbal responses of each of the two reactor operators. The data gathered by the

behavior logs successfully served several functions. First, the logs served as a way to integrate the simulator logged data as well as the video recordings, conducted by yet another human factors expert. This report does not include analysis of the video recording or simulator logs. However, in order to effectively examine operator performance in the future, the behavioral data must be incorporated with the simulator logs in order to capture the operators' interaction with the turbine control system. The behavioral logs capture the actions of the operators, while the simulator log captures the change in the system states resulting from the operators' actions. In addition to documenting the operators' actions, the logs serve as a way to examine the rationale for those behaviors as the operators' vocalization of indicator levels and actions provide insight into the mental model the operators have for the system. The behaviors and rationale for particular actions will serve as important benchmarks for evaluating operator performance and usability in the new DCS before it is placed in service at the plant.

4.2.2.2 Simulator Logs and Video Recordings

The plant simulator logs all actions taken by the operators (e.g., actuation of a switch on the board), all events initiated by the instructor (e.g., small steam generator leak), all alarms received, and select plant parameters (e.g., turbine speed). These logs are time-stamped and can be synchronized and integrated with other data such as behavioral logs. The simulator logs allow us to quantify the time that operators spend completing particular steps of the procedures associated with the turbine control system. The simulator logs and video recordings are invaluable for current and new system benchmarks such as comparing the time to complete tasks, the amount of information conveyed between the shift crew for a particular task, the rate of information throughput between operators, the number of actions required to complete a task, and the number of operator errors committed during a task.

4.2.2.3 Structured Discussions

At the end of each scenario walkthrough and also at key transition points throughout the workshop, an independent human factors consultant led a high level discussion concerning any potential issues. The goal of this portion of the workshop was to maintain a high level perspective of the proposed turbine control system and to ensure that key parameters were visible to the operators and key controls were appropriately placed throughout the HSI so that operators could quickly access them during time sensitive evolutions. One of the key findings from these structured discussions was the lack of need the operators conveyed for several parameters that were previously thought to be critical for scenario walkthroughs. With this knowledge, the unnecessary parameters can be replaced by something the operators deem more useful for turbine control.

4.2.2.4 Semi-Structured Discussions

During the majority of the workshop, with the exception of the four scenario walkthroughs, the operators were guided by the human factors experts to ensure that the discussion covered all the relevant usability issues. The unique constituency of this workshop allowed the discussion to rapidly identify any problematic issues, clarify how each aspect of the system was designed to be operated, and also gain feedback from the operators about how they would operate the system on a daily basis. There were several instances, in which the operators and design team deviated from the structured guidance and discussed concept of operations and physical system components and controls. During these unstructured discussions a number of important findings emerged. These centered on the concept of how the HSI was designed to operate versus how it would actually be operated by the team of operators.

4.2.3 Advanced Measures

Each operator completed a set of the three forms after completing each turbine control related scenario for both the existing and the new prototype TCS.

4.2.3.1 Situation Awareness

A modified version of the Situation Awareness Rating Technique (SART) originally developed by Taylor (1990) to assess aircrew situation awareness was provided to the operators as a measure of situation awareness (SA) after each of the workshop scenarios were completed by the crew of operators. The SART was developed through a series of interviews with pilots concerning the role of SA in flight scenarios. First, a large set of flight scenarios involving SA in general sense were generated from veteran pilots. Then, a separate set of pilots examined subsets of these flight scenarios consisting of groups of three scenarios. The pilots selected one of the three scenarios that was meaningfully different from the others in terms of SA demands. The researchers then compiled a set of SA constructs consisting of the pilots' reported rationale to account for the perceived differences in SA. The researchers conducted a principal component analysis on this set of SA constructs to yield 10 generic SA constructs that became the 10 dimensions of the SART. Each of the dimensions provides scale in which the participant can provide a rating for their subjective experience within that dimension during the scenario.

A modified version of the SART (see Table 5) was selected to assess operator SA due to its speed of administration and simplicity. The SART requires only a few moments to administer after the completion of each trial. Quick administration is important because it reduces any memory requirements for the participants, such that the participant can accurately recall and report subject perceptions experienced during the scenarios. Additionally, the quick administration is advantageous because it reduces the time demand placed on the operators and support personnel providing expertise during the workshop. Any means to reduce the amount of time required to gather data on operators reduces study costs and allows that time to be allocated efficiently for other critical needs. The simplicity of the SART is also advantageous because it elicits easily understood SA construct ratings from the operators and affords easy interpretation of the ratings following the study administration. Though the SART is relatively intuitive, the language was modified to improve the understandability by operators as well as reflect the language used within the nuclear process control domain.

The modified SART included the nine dimensions of SA with a rating scale provided for each dimension as depicted in Table 6. The operators were provided with a paper form containing the nine SA dimensions along with an anchor bar scale for each dimension. The scale ranged from one through ten with the appropriate anchor word placed at each pole of the scale. All dimensions were worded neutrally and the scales were laid out with low magnitude anchor word descriptors on the left pole and high value anchor descriptors on the right. This layout was used to maintain consistent positive and negative magnitude directions to prevent any participant confusion experienced while completing the form.

The modified SART is scored in terms of the overall SA demands generated by a situation. The lower the SART score, the lower the SA demands. In this particular context, the interface design drives the SA demands. Ideally, the new TCS should result in a lower modified SART score than the existing TCS interface.

Table 5. The Modified Situation Awareness Rating Technique.

DATE: ___ /___ /___ Scenario: _____

ID: ___

SART

Stable or Rapidly Changing Scenario

How changing is the scenario? Is the scenario highly dynamic and likely to change suddenly (High) or is it very stable and straightforward (Low)?

Low High

Simple or Complex Scenario

How complicated is the scenario? Is it complex with many interrelated components (High) or is it simple and straightforward (Low)?

Low High

Few or Many Factors Changing During the Scenario

How many variables are changing within the scenario? Are there a large number of factors varying (High) or are there very few variables changing (Low)?

Low High

Level of Alertness

How engaged are you by the scenario? Are you alert and ready for activity (High) or do you have a low degree of alertness (Low)?

Low High

Attention Required

How much are you concentrating during the scenario? Is your attention in high demand during the scenario (High) or in low demand (Low)?

Low High

Division of Attention

How much is your attention divided by the scenario? Are you concentrating on many aspects of the scenario (High) or focused on only one (Low)?

Low High

Mental Workload

How much mental workload do you have to spare during the scenario? Do you have sufficient time and memory to attend to the variables (High) or insufficient time and memory (Low)?

Low High

Amount of Information

How much needed information have you gained during the scenario? Was all the information you needed available and understood (High) or some missing and not understood (Low)?

Low High

Familiarity with Scenario

How familiar are you with the scenario? Do you have a great deal of relevant experience (High) or is it a new scenario (Low)?

Low High

Table 6. Situation Awareness Dimensions.

Dimension	Continuum
Stability	Rapid - Stable
Complexity	Simple - Complex
Changing Factors	Few - Many
Alertness Level	Low - High
Attention Required	Low - High
Division of Attention Focus	Single – Many
Mental Workload	Time/Memory*
Amount of Information	Low - High
Familiarity	Typical - Novel

4.2.3.2 NASA-TLX

The NASA Task Load Index (TLX) is probably the most common measure used to assess workload across a number of domains including nuclear process control. The NASA-TLX has undergone extensive development and serves as the gold standard for workload assessment within the field of HFE (Hart and Staveland, 1988; Hart, 2006). The NASA-TLX evaluates the following components associated with workload:

- Mental Demand
- Physical Demand
- Temporal Demand
- Performance
- Effort
- Frustration Level

Operators were provided with a paper-based version of the NASA-TLX. Each of the dimensions included a 10-point rating scale similar to the scale included with the modified SART form. However, the NASA-TLX dimensions all use the same low to high scale magnitudes and anchor descriptors. Table 7 contains the paper-based copy of the NASA-TLX provided to the operators.

4.2.3.3 Performance Shaping Factors

Performance shaping factors are aspects of a situation that can either aid or hinder performance. Twelve performance shaping factors based on those used for operator studies on human reliability analysis (Forester et al., 2014) were included in the paper-based measure provided to the operators. The operators were tasked with indicating whether each of the factors was positive, i.e. improved performance, or negative, i.e. decreased performance during each scenario. The paper-based form provided to the operators can be seen in Table 8.

Table 7. The NASA-TLX Used to Assess Workload.

DATE: ___ /___ /___ Scenario: _____

ID: ___

NASA TLX

Mental Demand

How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)?

Low High

Physical Demand

How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)?

Low High

Temporal Demand

How much time pressure did you feel due to the rate or pace at which the simulation occurred?

Low High

Performance

How successful do you think you were in accomplishing the goals of the simulation?

Low High

Effort

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

Low High

Frustration Level

How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?

Low High

Table 8. The Performance Shaping Factor Worksheet.

DATE: ___ /___ /___ Scenario: _____
ID: ___

Performance Drivers

Identify if any of the following factors were drivers on your performance during the scenario. If you believe a factor improved your performance, please circle POSITIVE. If a factor decreased your performance, circle NEGATIVE. If it does not apply, circle N/A.

Adequacy of Time Did you have plenty of time (POSITIVE) or not enough time (NEGATIVE) to complete the tasks?	POSITIVE	NEGATIVE	N/A
Time Pressure Did your perception of time limits improve (POSITIVE) or hinder (NEGATIVE) your performance?	POSITIVE	NEGATIVE	N/A
Stress Did you feel stressed during the scenario? Did it improve (POSITIVE) or hinder (NEGATIVE) your performance?	POSITIVE	NEGATIVE	N/A
Scenario Complexity Was it easy (POSITIVE) or difficult (NEGATIVE) to diagnose the situation properly?	POSITIVE	NEGATIVE	N/A
Execution Complexity Was it easy (POSITIVE) or difficult (NEGATIVE) to carry out the activities in the scenario?	POSITIVE	NEGATIVE	N/A
Indications of Conditions Were the indications and alarms clear and helpful (POSITIVE) or unclear or misleading (NEGATIVE)?	POSITIVE	NEGATIVE	N/A
Human-Machine Interface Is the labeling and organization of information clear and helpful (POSITIVE) or unclear and misleading (NEGATIVE)?	POSITIVE	NEGATIVE	N/A
Training Did your previous knowledge of this type of scenario improve (POSITIVE) or hinder (NEGATIVE) your performance?	POSITIVE	NEGATIVE	N/A
Experience Did your previous practice on this type of scenario improve (POSITIVE) or hinder (NEGATIVE) your performance?	POSITIVE	NEGATIVE	N/A
Procedural Guidance Did the procedures adequately aid (POSITIVE) or fail to aid (NEGATIVE) the completion of the tasks?	POSITIVE	NEGATIVE	N/A
Communication Was needed information successfully exchanged (POSITIVE) or were there gaps in the information exchanged (NEGATIVE)?	POSITIVE	NEGATIVE	N/A
Team Dynamics Did the tasks lead to the crew working as a team (POSITIVE) or did the crew members need to work independently (NEGATIVE)?	POSITIVE	NEGATIVE	N/A

4.2.3.4 Results

The sample size for this example study was quite small (n=4), which prevented any formal statistical analysis from being performed. However examining the data from each measure revealed several trends with potentially significant design implications. These trends provide valuable diagnostic information, which serve to benchmark the new TCS against the existing TCS in support of the V&V process.

Situation Awareness

Overall, the modified SART score, which aggregates the SA dimensions into a single value representing the SA demands generated by the situation, was lower for the new TCS than the existing TCS. Lower demand scores correspond to increased SA. Differences in each dimension of SA were also observed in the modified SART data, but these differences were often quite small. The existing and new TCS HSIs differed by an average across participants of 0.13 and 0.17, for the familiarity and attention required dimensions, respectively. Since these ratings are based on a 10-point scale, a difference of 0.13 and 0.17 is quite small. Some of the dimensions differed more dramatically, such as the division of attention, complexity, and amount of information, with reductions in dimension scores of 1.17, 1.42, and 1.75 respectively for the new TCS. The success of the new TCS over the existing TCS is evident simply by looking at the largest reduction in SA demand dimension of amount of information. The lower SA demand reported by the operators for the new TCS suggests that the new layout eliminates the amount of information the operators previously had to remember when interacting with the TCS. The new system, which aggregates and synthesizes much of this information, does not require the operators to remember this information since it is displayed prominently near where it is required within the new TCS. The same pattern can be observed for the division of attention dimension. The division of attention SA demand was reduced by the new TCS due to this aggregation and synthesis of information. Even with such a small sample size and no statistical analysis, simple and diagnostic benchmark comparisons can easily be gleaned from the modified SART form. Even if a conservative interpretation of the modified SA scores were adopted, the trends observed for each dimension towards a reduction in SA demands suggests that the new system certainly does not decrease performance. These benchmark comparisons become integral evidence for the V&V process. To meet the NUREG-0711 requirements of demonstrating the new system meeting and exceeding the performance of the existing system, the modified SART serves as a good measure.

NASA-TLX

As with the modified SART, the NASA-TLX measure relied on a sample that was too small for formal statistical analysis. Overall, the NASA-TLX scores showed modest improvement for the new TCS HSI, with an average difference of 0.46 across all participants. The physical demands did not change between the existing and new interfaces, which reduced the overall difference captured by the NASA-TLX score. However, trends for two dimensions emerged from the NASA-TLX measure that demonstrate a substantial improvement of the new TCS over the existing TCS. The primary differences between the existing and new prototype TCS were observed for the mental demands and temporal demands dimensions as reported by the operators. The new prototype TCS demonstrated average reductions in mental and temporal demands of 0.92 and 1.17 respectively. The reductions along these dimensions suggest that the new TCS interface reduces the time and mental effort required to interact with the turbine. These two dimensions are likely the most important dimensions of the NASA-TLX as it relates to turbine control, since the operators previously had to mentally synthesize a large amount of information spread across numerous controls on the main control boards. The NASA-TLX data further bolsters the

modified SART data suggesting meaningful improvements of the new prototype TCS over the existing TCS. Concretely demonstrating these improvements is vital to the V&V process in order to adhere to regulatory requirements.

Performance Shaping Factors

The performance shaping factors did not differ between the existing and new prototype TCS. Ideally, an increase in positive ratings would be a desirable observation for the new prototype TCS; however, similar ratings are still considered a beneficial outcome for these benchmark tests. According to NUREG-0711, any changes to the system should not negatively impact operation of that system. As long as the new prototype TCS demonstrates comparable performance, the benchmark test is considered a success and the new prototype TCS can be classified as compliant with NRC regulations. Because the performance shaping factor scale in Table 8 is a compressed scale (only offering positive vs. negative influences), greater sensitivity to these changes may be possible by expanding the range of possible values operators can select. A revised scale with degrees of positive or negative influence will be developed for future applications.

4.2.3.5 Utility of Advanced Measures

The discussion on simple measures in Section 3.4 provided a caution on the use of advanced measures for V&V purposes. The overhead of administering some advanced measures may not provide a good value proposition, especially if those measures do not readily inform the V&V process. The advanced measures employed here attempted to maintain the advantages of simple measures. These measures were easily administered very quickly, which makes them ideal for conducting benchmark comparisons between existing systems and the new systems. Additionally, the measures are easily interpreted, which further reduces the time and effort required to complete new system benchmarking. Any costs that can be saved to the power utility during the V&V process should be pursued. Using simple measures like these help reduce the cost of conducting the necessary human factors studies required for the NUREG-0711 compliant V&V process.

4.3 Verification: Expert Review

4.3.1 NUREG-0700 Evaluation

The U.S. NRC's NUREG-0700, Rev. 2, *Human-System Interface Design Review Guidelines* (O'Hara et al., 2002) provides an exhaustive set of guidelines for reviewing the human factors aspects of HSIs for nuclear power. The guidelines are organized into four parts:

- Part I addresses basic HSI elements.
- Part II addresses specific systems like alarm systems and computer-based procedures.
- Part III documents ergonomic considerations.
- Lastly, Part IV focuses on the maintainability of digital systems.

Over the entirety of the document there are 2,195 guidelines. Because the TCS interface is still under development, a full NUREG-0700 is not warranted nor desired. Each of the four parts contains additional hierarchical levels that help in managing the complexity of the guidelines. Several sections are not applicable to digital interfaces, some sections were simply outside the scope of this workshop (e.g. workplace design), and other sections just could not be properly evaluated until the interface is closer to its final form. To systematically select an applicable, practical, and informative subset of guidelines, three human factors experts sorted through all 2,195 guidelines and selected a set of guidelines they thought should be included in the evaluation using an internally generated spreadsheet version of the NUREG-0700. After initial selections were formed by the individual analysts, a set analysis was conducted to identify agreement among the selections. On the first iteration there were 27 criteria selected by all three human factors experts and 69 additional criteria selected by two of the three human factors experts. Each human factors expert then examined the set difference between the union of their peers and their own selections. In the second round of selections 43 criteria were common to all three human factors experts. From the second set analysis the most senior human factors expert selected the final set of guidelines. After this selection we were able to pare the 2,195 criteria to 78 factors.

4.3.2 Heuristic Evaluation

NUREG-0700 could be said to be the gold standard for evaluating HSIs relating to nuclear power. However, its comprehensiveness comes at a cost, and the conclusions drawn from a NUREG-0700 evaluation are more appropriate for a summative evaluation than a formative evaluation. For the formative evaluation conducted as part of this study, our goal is simply to capture expert impressions and provide organized, directed, precise, and unambiguous feedback to the interface designers. NUREG-0700 makes this difficult because the shear volume of guidelines can obscure the most relevant details. Furthermore, the language and jargon used in NUREG-0700 is not entirely accessible to non-HFE practitioners. While planning the workshop we knew we wanted a heuristic evaluation that we could use to assess the new TCS interface screen-by-screen. We also wanted to be able to use the evaluation to conduct an expert review as well to provide guidance in conducting a group review with operators during the workshop. With dozens of screens to evaluate, whatever checklist we decided to use needed to be simple and straightforward. For these reasons we quickly steered away from NUREG-0700 as a primary expert evaluation tool.

Our second thought was to revert to Nielsen's Usability Heuristics (1993) and Gerhardt-Powal's Cognitive Engineering Principles (1996). Within the LWRS evaluation team, we had prior experience with this approach (Ulrich et al., 2012). While we could agree that the Nielsen and Gerhardt-Powal indices were better suited to the task at hand than NUREG-0700, we still found them less than ideal. The principles and heuristics in both tended to be too open-ended and, like NUREG-0700, used language that might not be accessible to operators. The end result is we opted to create a tailored checklist based on the most relevant items from NUREG-0700, Nielsen's Usability Heuristics (1993), and Gerhardt-Powal's Cognitive Engineering Principles (1996). Because our evaluation was constrained to a DCS for nuclear power, our resulting checklist could revise the domain independent language from Nielsen and Gerhardt-Powal. The checklist was also made less ambiguous by phrasing the modified checklist as questions so that they could be answered "Yes" or "No." Further optimizations were made, by creating one checklist that could be applied holistically across the interface, and a second set to evaluate individual screens. The checklists are provided below in Table 9 and Table 10, respectively.

Table 9. Holistic Checklist.

Navigation

- Are the methods of navigation clear and consistent?
- Is the sequence of screens logical?
- Can the user navigate readily from page to page?
- Is it easy to return to the supervisory screen?

General Interface

- Is the use of titles and acronyms consistent across pages?
- Is the use of controls and indicators consistent across pages?
- Is the use of color consistent across pages?
- Are there missing pages?
- Are there pages that could be combined?
- Is critical information available across the interface?
- Can the interface be simplified?

Alarms

- Is the general embedded alarm philosophy acceptable?
- Is the alarm highlight meaning clear on individual indicators?
- Is the navigation to alarming components appropriate?

Table 10. Screen-by-Screen Checklist.

Content and Objectives

- Is anything unclear on the page?
- Is the screen title consistent with the content?
- Do the indicators and soft controls contribute to the purpose of the page?
- Are relevant indicators or soft controls missing?
- Are there unnecessary indicators or soft controls?
- Are unwanted tasks automated (mental calculations, estimations, comparisons, and unnecessary thinking)?

Screen Layout

- Is the screen easy to read for the user?
- Are the screen elements logically organized?
- Is the screen layout consistent?
- Does screen element organization support understanding the process?
- Are characters and symbols readable, attractive and properly sized?
- Is the information density on the screen appropriate?
- Does the use of color aid the user in understanding the content?
- Can the screen be simplified?

Indication and Control

- Is the meaning and state of each indicator clear?
- Is the function and state of each soft control clear?
- Is the relationship between controls and indicators clear?
- Are the labels for controls and indicators appropriate and consistent?

4.3.3 Style Guide Adherence

In addition to NUREG-0700 and the heuristic evaluation, the human factors experts identified and reported discrepancies between the plant's existing HSI style guide and the new TCS interface. Due to the proprietary nature of the style guide, the nature and results of this review will not be discussed in further detail.

4.4 Summary of Findings

The evaluation processes described above yielded a significant number of potential usability issues and suggested changes to improve the usability of the HSI. The changes comprised a variety of issues, but the most prominent varieties included labeling nomenclature, task-oriented screen layout, alarm structure, and navigation. For each issue, a consensus was drawn from the operators to resolve potential fixes. The changes were cataloged and directly provided to the HSI design vendor for incorporation into the next iteration of the turbine control HSI. Issues tended to be discovered by several of the evaluation methods and measures. Here we have corroborated and organized the discovered issues.

4.4.1 Nomenclature

Due to the number of different entities with specialized domain expertise a variety of issues related to the nomenclature emerged. The majority of the issues were simple differences in the terms and acronyms commonly used by the operators but unknown to the rest of the design personnel. For example, the acronym OP standardly refers to an "operating procedure" while the interface designers used the term OP refers to "output." Turbine control engineers use the acronym "PID" to describe a specific type of controller (Proportional, Integrative, Derivative controller) while the operators' intuitions suggested PID referred to "Piping and Instrumentation Diagrams." See Figure 4 for an example. The workshop identified several of these discrepancies and the agreed nomenclature was provided to the HSI software designers (see Table 11).

4.4.2 Indication and Control

Another recommendation was to add clearly identifiable titles to convey the automatic versus manual modes of operation and their corresponding indicators and controls. For example, in Figure 4, the Raise and Lower controls are only available when the speed controller is in Mannual Mode. The Ramp Rate Selection and Ramp Rate Entries are only available when the controller is in Auto Mode. These mode-specific settings were not clear to operators in the workshop.

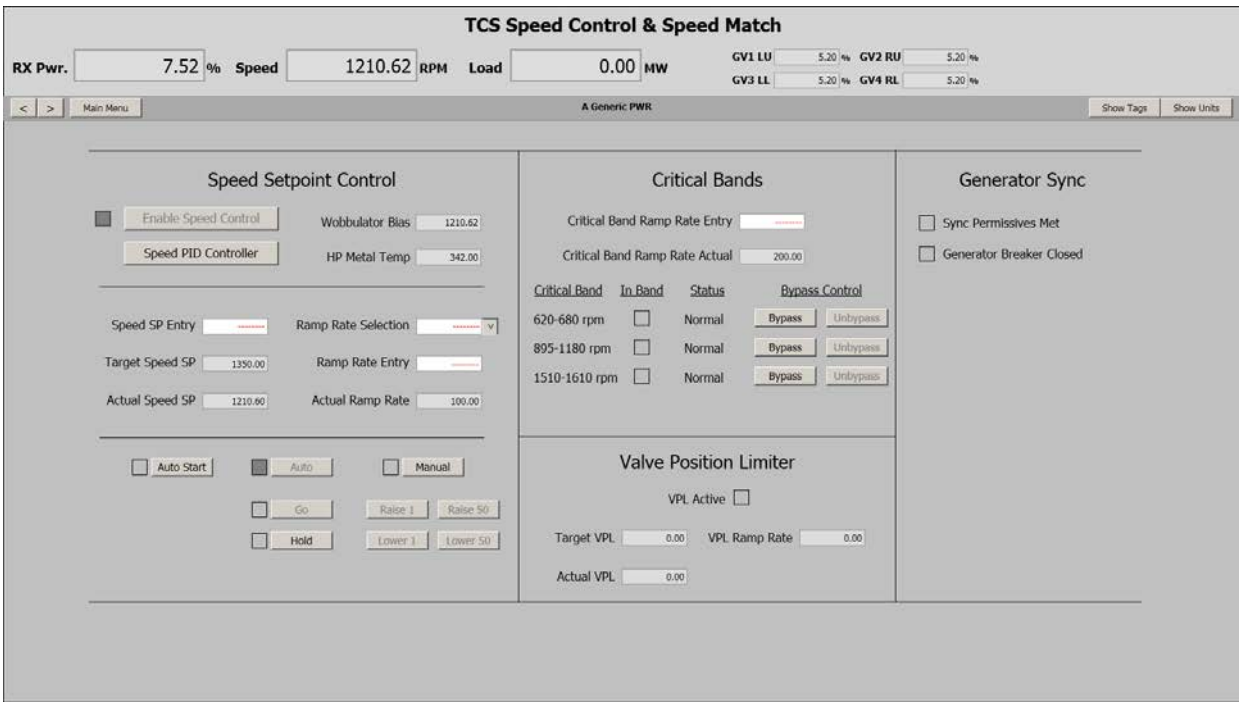

Figure 4. A Speed Control Interface Screen.

Table 11. Suggested Changes in Control Interface Nomenclature.

Term	Explanation	Suggested Revision
OP	"OP" means "Operating Procedure" to operators.	"Output" or remove
PID	"PID" could mean "Piping and instrumentation diagram."	"Controller"
SP	Setpoint or the demand value of the controller.	remove
Target SP	Known as "demand" or "setter" to operators.	"Target"
Actual SP	Conveys the current setpoint of the controller. During the go phase this value changes from the starting value to the target value. Once the target value is reached Actual SP is the Target SP. Operators do not currently have this value available.	"Demand"
GVPC	Intended to mean "Governor Valve Position Control." Operators are familiar with "Impulse in."	"Impulse in"
Auto	Known as "Oper Auto."	"Oper Auto" or "Auto"

4.4.3 Organization and Navigation

Another group of potential issues consisted of task-oriented screen layout. The digital control eliminates the current physical controls by creating digital replicas in a series of screens comprising the DCS. As with the physical control boards, real estate for individual indicators and controls is limited due to the size constraints of the LCD (liquid crystal display) and the shear number of elements that must be displayed. Due to the real estate limitation, organizing the indicators and controls requires a large number of screens, each organized based on a particular turbine control task. Designing the screens so that all the relevant indicators and controls are present for a given task requires a considerable amount of effort to match the physical system with the operator's mental model. The workshop revealed a number of potential issues in which extraneous indicators and controls were included in a screen. The operators simply expressed a lack of need for these items. Another prevalent result concerning the task oriented screen layout, was the identification of missing indicators and controls the operators indicated they would use for the particular task associated with the screen. Governor valve positions were revealed to be of particular importance for a variety of tasks. The governor valves were not important enough to merit inclusion in a designated area in all of the screens, but rather the operators expressed interest in the governor valves being present in several key screens. To maintain consistency and allow the operators to quickly assess the governor valve positions, the human factors experts suggested the same orientation and presentation of the governor valve indicators within the same location of each screen across the different screens in which the governor valves appeared.

Figure 4 illustrates how this might be accomplished by placing these indications in the upper right of the display area. Furthermore, some screens supported multiple tasks associated with different functional aspects of turbine control, however the operators expressed confusion since the grouping of the tasks within a single screen did not match their mental model, at least in accordance with the current operation practices for turbine control.

Navigation issues were also identified during the expert review. The navigation scheme uses a main navigation menu that allows the operators to move to a particular subsection of the turbine control system. The primary issue identified with the navigation is the use of arrow buttons to toggle between levels of the system and toggle horizontally within the same level of the system to related components and the tasks associated with those components. As the operators' mental models become more mature, the arrow buttons may prove efficient for moving between screens, but currently they afford no indication as to what screen they will present. During the workshop, operators were never observed to use the arrow navigation buttons. They always navigated back to the Main Menu before selecting the desired screen. Providing dynamic labels on the arrow buttons was recommended to aid the operators in orienting themselves within the system. Operators can also cue additional status information for particular components; however in the majority of cases the operators expressed little interest in viewing this extra information since in most cases the additional information was unnecessary for their task. The human factors experts recommended that the HSI design team eliminate the majority of these additional information navigation links based on the operators' feedback for specific instances of unnecessary additional information.

4.4.4 Alarm Presentation and Management

Since an operator must select an individual screen for viewing while completing a turbine control task, the operator must have the ability to leave a screen and access alarm information for components not

represented in the currently viewed screen. The alarm presentation philosophy was extensively examined during the expert review. In the event of an alarm, the operators were guided to navigate to additional alarm information from at least one location within any given screen. Some navigation menu screens offer additional methods for quickly navigating to the alarmed component; however, the majority of interactions the operators have with alarms consist of viewing and manipulating alarmed components in an alarm list based primarily on the chronological order each alarm is triggered. The operators can filter the alarms to help overcome alarm cascades, but the system does have some potential for alarm flooding.

Unfortunately, there are limitations to ensuring the visibility of the alarms, since the screens would become overly complex and cluttered. Fortunately, the operators expressed confidence in working with the alarm lists. Priority filters, which were incorporated but not functional at the time of the workshop, were emphasized by the human factors experts as a critical method for isolating the key alarms during the time critical diagnostics steps required by a plant malfunction. Specific system filters that limit alarms to only a given system and level of system to be displayed were recommended as default filters for the turbine control system.

4.4.5 Ergonomic Considerations

Due to logistics at the plant, one of the two turbine control stations will be implemented without a physical keyboard or physical numeric entry device. Operators will have a trackpad but no other means of inputting information. Some of the controls require operators to input a numeric target setpoint or ramp rate. The workshop discussed several possible alternatives such as finding room for a numeric keypad, providing an on-screen numeric keypad, providing drop-down selection dialogues, or possibly not using this keyboard-less terminal for tasks requiring numeric entry.

A second ergonomic issue due primarily to space constraints was identified. In the planned panel retrofit, two 1080p LCD displays are intended to be placed on the control board and serve as a primary means of displaying the new DCS. Due to space constraints, these displays are smaller than typical for their pixel resolution and have a pixel density of 115 dpi compared to a more typical density of 72 dpi. This fact combined with an unusually deep control board results in the operators experiencing difficulty reading particular screen elements that are not designed to be scalable. In particular, the DCS Alarm Summary page lists alarms in a table with 10 point font. When dot pitch and viewing distance are taken into consideration, the resulting viewing angle of 10 point font is only 8.2 minutes of arc. NUREG-0700 conservatively specifies a minimum of 16 minutes of arc for font size. Perhaps counterintuitively, this issue could be addressed by using displays with lower pixel density (lower resolution, higher dot pitch) or by identifying a means of scaling the presented information.

5. CONCLUSIONS

5.1 Usability Measures

This report has provided a snapshot of human performance measures and methods suitable for application during the design phase of control room modernization. It first provided background on the use of formative methods for verification and validation, next overviewed human performance measures that can be obtained for control room validation, and concluded with an example of a study combining formative verification and validation to evaluate a proposed digital TCS upgrade for the control room at an existing TCS.

Collectively, the various measures effectively captured a large number of potential human factors usability issues for the proposed TCS. The workshop data collection successfully achieved its primary purpose to quickly, concisely, and unambiguously provide the HSI design vendor with specific recommendations to improve the next iteration of the TCS.

In addition to high level results from the data collection method, there is a large amount of data that requires further analysis. The simulator logs, video recordings, and behavioral logs comprise a rich data set that will prove invaluable during the next phase of the project in which a more direct comparison between the current turbine control operations and the new TCS will be possible. For the formative evaluation of the static displays, the behavior logs, structured and semi-structured discussions, and expert reviews provided the right quality and level of information needed to inform the next phase of the TCS design.

Though the primary purpose of the workshop and data collection method was not to validate the HSSL as an effective tool for prototyping and usability studies, clear evidence emerged based on the operators' feedback that the HSSL did in fact admirably perform well and that operators felt the glasstop representation of the plants' control boards was accurate and usable.

It is anticipated that as this project continues to support the TCS modernization across additional plants at the utility, we will produce additional reports to further elucidate human performance measures for control room modernization. As depicted earlier in Figure 2, this report only addresses evaluation of displays up to the 70% TCS completion point. At this stage, only a subset of human performance measures is applicable. As the design of the TCS is finalized, it will become possible to employ summative measures appropriate for ISV. As the TCS is implemented across the fleet of reactors at this utility, it will be possible to provide guidance to demonstrate the process of applying a design across a fleet and determine where design and evaluation efficiencies are possible through reuse and standardization. Finally, additional measures beyond those obtainable in a standard control room configuration—such as eye tracking and physiological measures—may be explored to arrive at additional guidance to aid utilities in their control room modernization activities.

5.2 Measures vs. Process

With this report, we wish to stress that there is no single set of measures that is optimal to helping perform V&V in support of control room modernization. Rather, there is a process that we have outlined, which can help ensure that the design of upgrades adheres to a trajectory of refining the HSI beginning early in the design phase. Through formative V&V, the process ensures that there are no human error deficiencies identified during final, summative V&V efforts. The process prescribed can be accomplished

using straightforward usability measures of operator performance—firstmost operator preference, but also in simple measures of workload, situation awareness, and performance drivers. These measures will prove sufficient to identify deficiencies in the design and to document improvements over time in the design.

It is accepted in the human factors community that it is better to be involved early in the design of a system rather than later (International Standards Organisation, 2010). This stems from the best window in the design cycle for our field to affect change. Change early in the design cycle—in the formative stages of system design—allows for the incorporation of user input to improve the design. Conversely, performing an assessment of a design late in the design stage—at the summative stage—risks finding fault in a nearly deployed system. Late-stage V&V hardly endears us as contributors to the end product, nor does it allow adequate time to fix issues that may surface in the system. As such, we advocate performing extensive V&V activities formatively, with summative evaluation simply replicating earlier V&V findings.

There is considerable value in conducting V&V across the design life cycle of the HSI. The key point here is that in the nuclear community, with its strong emphasis on summative evaluation in the form of ISV, we potentially put ourselves in the position of doing human factors at the tail end of the design process, when we are relatively speaking least able to improve the design. There is nothing prescribing this tendency toward late-stage evaluation. It may be a simple confusion over the guidance in NUREG-0711, which is foremost a document guiding regulatory review at the completion of the design cycle rather than an exhaustive best practice for human factors. The propensity for late-stage V&V may also be a result of a certain disclosure hesitancy between the licensee and the regulator, in which the intermediate steps of a design—the designs with shortcomings that might be revealed through operator studies—are not readily shared as part of a license submission. The problem is that when V&V is relegated to a tail-end activity, we have not necessarily engaged in a process of system improvement based on user input and evaluation. Nor have we documented lessons learned in the design process. We tend to focus on demonstrating that the overall system as designed actually worked. We haven't demonstrated that the design evolved to the point of working. We seek to rubber stamp design rather than actively refine it.

We suggest that the human factors community needs to reassert V&V not just as ISV but also as part of an iterative user centered design process. Experience in other domains—e.g., educational testing, safety cases, and quality control—reveals the advantages of early and frequent sampling of progress to demonstrate a successful process. We need to understand and document stumbling blocks that weren't good design ideas. These ideas need to be shared by licensees as welcome byproducts of the design process. Equally importantly, design foibles that are overcome through early-stage and iterative V&V should be championed by regulators as artifacts of an effective human factors process.

In short, for human factors to be truly effective for nuclear applications, there needs to be a shift from late-stage ISV to early-stage V&V. This in not to downplay the importance of ISV; rather, it is to ensure that human factors can help shape and optimize the design of the HSI leading up to ISV. ISV is the culmination of earlier human factors efforts, not a substitute for them.

6. REFERENCES

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

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

Boring, R.L., Hendrickson, S.M.L., Forester, J.A., Tran, T.Q., and Lois, E. (2010). Issues in benchmarking human reliability analysis methods: A literature review. *Reliability Analysis and System Safety, 95,* 591-605.

Boring, R., Joe, J., & Ulrich, T. (2014). *Strategy for Migration of Traditional to Hybrid Control Boards in a Nuclear Power Plant, INL/EXT-14-32534.* Idaho Falls: Idaho National Laboratory.

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

Dumas, J.S. and Redish, J.C. (1999/1993) *A Practical Guide to Usability Testing.* Bristol, UK: Intellect.

Electrical Power Research Institute. (2005). *Human Factors Guidance for Control Room and Digital Human-System Interface Design and Modification, 1010042*. Palo Alto: Electrical Power Research Institute.

Forest, J., Dang, V.N., Bye, A., Lois, E., Massaiu, S., Broberg, H., Braarud, P.Ø., Boring, R., Männistö, Liao, H., Julius, J., Parry, G., and Nelson, P. (2014). *The International HRA Empirical Study: Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data, NUREG-2127.* Washington, DC: U.S. Nuclear Regulatory Commission

Fuld, R.B. (2007). On system validity, quasi-experiments, and safety: A critique of NUREG/CR-6393. *International Journal of Risk Assessment and Management, 7,* 367-381.

Gerhardt-Powals, J. (1996). Cognitive engineering principles for enhancing human-computer performance. *International Journal of Human-Computer Interaction, 8,* 189-221.

Hamblin, C., Cataneda, M., Fuld, R.B., Holden, K., Whitmore, M., and Wilkinson, C. (2013). Verification and validation: Human factors requirements and performance evaluation. *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting,* 2032-2036.

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

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

Hugo, J., Boring, R., Hanes, L., and Thomas, K. (2013). *A Reference Plan for Control Room Modernization: Planning and Analysis Phase, INL/EXT-13-30109.* Idaho Falls: Idaho National Laboratory.

International Standards Organization. (2010). *Ergonomics of Human-System Interaction—Part 210: Human Centered Design for Interactive Systems, ISO 9241-210.* Geneva: International Standards Organization.

Lew, R., Boring, R.L., & Ulrich, T.A. (2014). A prototyping environment for research on human-machine interfaces in process control: Use of Microsoft WPF for microworld and distributed control system development. *Proceedings of the International Symposium on Resilient Control Systems (Resilience Week).*

Mirowski, P. (1999). "The Ironies of Physics Envy" in *More Heat Than Light*. Cambridge, UK: Cambridge University Press.

Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. *Computer-Human Interaction (CHI) Conference Proceedings* (pp. 152-158). New York City: Association for Computing Machinery.

O'Hara, J.M., Brown, W.S., Lewis, P.M., and Persensky, J.J. (2002). *Human-System Interface Design Review Guidelines, NUREG-0700r2*. Washington, DC: U. S. Nuclear Regulatory Committee.

O'Hara, J.M., Higgins, J.C., Fleger, S.A., and Pieringer, P.A. (2012). *Human Factors Engineering Program Review Model, NUREG-0711, Rev. 3.* Washington, DC: U.S. Nuclear Regulatory Commission.

O'Hara, J., Stubler, W., Higgins, J., and Brown, W. (1995). *Integrated System Validation: Methodology and Review Criteria, NUREG/CR-6393.* Washington, DC: U.S. Nuclear Regulatory Commission.

Redish, J., Bias, R.G., Bailey, R., Molich, R., Dumas, J., and Spool, J.M. (2002). Usability in practice: Formative usability evaluations—Evolution and revolution. *Proceedings of the Human Factors in Computing Systems Conference (CHI 2002)*, 885-890.

Rubin, J. (1994). Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. New York, NY: John Wiley & Sons.

Scriven, Michael (1967). The methodology of evaluation. In Stake, R. E. (ed.), *Curriculum Evaluation*. Chicago: Rand McNally.

Taylor, R. M. (1990). Situational awareness rating technique (SART): the development of a tool for aircrew systems design. *Situational Awareness in Aerospace Operations 3*, 1-17. NATO-AGARD, Neuilly Sur Seine, France.

Tullis, T. & Albert, W. (2008). *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Elsevier/Morgan Kaufmann, Burlington, MA.

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

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

Vicente, K. J. (1997). Should an interface always match the operator's mental model? *CSE-RIAC Gateway, 8*, 1-5.