INL/EXT-15-36704

Light Water Reactor Sustainability Program

Verification and Validation of Digitally Upgraded Control Rooms

September 2015

U.S. Department of Energy

Office of Nuclear Energy

DISCLAIMER

This information was prepared as an account of work sponsored by an agency of the U.S. Government. Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness, of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.

INL/EXT-15-36704

Verification and Validation of Digitally Upgraded Control Rooms

Ronald Boring[1] and Nathan Lau[2]

[1]Idaho National Laboratory
[2]Virginia Polytechnic Institute and State University

September 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 U.S. commercial nuclear power utilities undertake significant control room modernization activities, they are adopting the human factors approach espoused in the U.S. Nuclear Regulatory Commission's NUREG-0711, Human Factors Engineering Program Review Model. Human factors serves to provide key design recommendations and confirmation that operator performance meets safety guidelines. NUREG-0711 outlines four broad phases for human-centered design activities and includes extensive discussion of the Verification and Validation (V&V) phase. In this report, we explore key aspects of V&V. We outline how V&V can be applied successfully across the design lifecycle, not just in support of late-stage integrated system validation. We particularly emphasize the benefits of early-stage V&V activities. We further discuss different types of evaluation, highlighting that there are multiple types of data that can inform and confirm a safe design, from operator-in-the-loop validation studies, to verification by experts against human factors standards, to knowledge transfer by expert users to the design team. We highlight the need for different data quality depending on the design phase and introduce the concept of As Low As Reasonable Assessment (ALARA) to apply discount usability evaluation principles to the control room design process. We suggest the safety case, adopted in many regulated safety-critical domains, as a framework for synthesizing different types of safety data. The report presents the Guideline for Operational Nuclear Usability and Knowledge Elicitation (GONUKE), an approach that focuses safety evaluation across design phases. Finally, we explore opportunities for future research on V&V. While the regulatory framework for system design demands conclusive results on operator performance through evaluation, the methods may not always provide the degree of conclusiveness that is needed. On-going V&V research is necessary to arrive at practical and defensible methods to evaluate operator performance while using upgraded systems.

(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. Several individuals have proved instrumental in bringing this body of research to fruition, including Thomas Ulrich, Roger Lew, and Kateryna Savchenko of the University of Idaho and Jeffrey Joe, Kirk Fitzgerald, Brandon Rice, and Heather Medema of INL.

(This page intentionally left blank)

CONTENTS

FIGURES

TABLES

ACRONYMS

ALARA	As Low As Reasonably Achievable/As Low as Reasonable Assessment
ALARP	As Low As Reasonably Practicable
CSNI	Committee on the Safety of Nuclear Installations
DCS	Distributed Control System
DOE	Department of Energy
EPRI	Electric Power Research Institute
FAT	Factory Acceptance Test
GONUKE	Guideline for Operational Nuclear Usability and Knowledge Elicitation
HED	Human Engineering Discrepancies
HFE	human factors engineering
HRA	Human Reliability Analysis
HSI	human-system interface
HSSL	Human Systems Simulation Laboratory
HUPESS	Human Performance Evaluation Support System
I&C	instrumentation and controls
INL	Idaho National Laboratory
ISO	International Standards Organization
ISV	integrated system validation
LTO	Long-Term Operations
LWRS	Light Water Reactor Sustainability
MCR	Main Control Room
NASA	National Aeronautics and Space Administration
NEA	Nuclear Energy Agency
NRC	U.S. Nuclear Regulatory Commission
NUREG	Nuclear Regulatory Document
OECD	Office of Economic Corporation And Development
OPAS	Operator Performance Assessment System
PRA	Probabilistic Risk Assessment
SA	situation awareness
SACRI	Situation Awareness Control Room Inventory
SAGAT	Situation Awareness Global Assessment Technique
SCED	single-case experimental designs
SHRP2	Strategic Highway Research Program 2
TLX	Task-Load Index
UK	United Kingdom
U.S.	United States
V&V	verification and validation
WGHOF	Working Group on Human and Organizational Factors

(This page intentionally left blank)

1. INTRODUCTION

Nuclear power plants in the United States (U.S.) have to date predominantly legacy control rooms comprised of analog or mechanical instrumentation and controls (I&C). Yet, new digital control systems and displays are readily available and have been extensively implemented in other process control industries (Hollifield et al., 2008; Strohbar, 2014). As noted across several previous reports (Boring and Joe, 2014; Boring et al., 2014; Boring et al., 2013), barriers to control room upgrades in the nuclear industry are multifold—from regulatory, to know-how, to plant downtime, to cost. Despite such barriers, there is a desire on behalf of many plants to move forward with control room modernization (Joe et al., 2012). Reliability issues of aging I&C, the cost of maintaining obsolete systems, training requirements for new operators, and successful international examples of control room modernization make a compelling case for upgrades, slowly overriding barriers.

The U.S. Department of Energy's (DOE's) Light Water Reactor Sustainability (LWRS) Program, together with the Electric Power Research Institute (ERPI) Long-Term Operations (LTO) Program, has championed research to assist the U.S. nuclear industry achieve control room modernization milestones. These milestones consist of proof-of-concept demonstrations of the upgrade process (Boring et al., 2014; Hugo et al., 2013; Ulrich et al., 2014) to serve as templates for the nuclear industry to proceed with upgrades as needed. The Human Systems Simulation Laboratory (HSSL; Boring et al., 2012 and 2013) at Idaho National Laboratory (INL) serves as a testbed for operator-in-the-loop studies on modernization, where nine simulator studies have been carried out to date using licensed reactor operators testing new digital human-system interfaces (HSIs) as part of industry upgrades. The findings from these modernization studies have been captured in both proprietary reports for the utility partners and summary reports to benefit the entire industry.

A key lesson learned from these operator-in-the-loop studies is the importance of human factors engineering (HFE) to the overall project outcome. As noted, the chief barrier to upgrades is not technological, and I&C replacement and upgrades are not simply an engineering problem. Rather, the systems being upgraded should result in improvements. Improvements may be marked by engineering metrics like reliability, but legacy systems have proved exceptionally reliable during their useful lifespan. In fact, while digital I&C may improve reliability compared to its analog antecedents, its effective lifespan may prove considerably shorter. So, digital technology may introduce the need for more frequent and costly replacement to maintain its performance advantage. The performance improvements from digital upgrades in the control room are likely to be found in terms of the reactor operators. Well-designed digital islands in the control room promise shorter training cycles, increased operator situational awareness, shorter response times in the face of plant upsets, and decreased human error. These goals cannot be achieved without careful consideration of the reactor operators. HFE serves as the bridge between technological solutions in the control room and the operators of those solutions. A poorly engineered solution is unlikely to yield significant operator performance improvements; in fact, it may actually introduce new error traps and decrease operational efficiencies.

The hallmark of HFE in the nuclear industry is compliance with applicable guidelines, particularly regulatory guidance such as NUREG-0711, *Human Factors Engineering Program*

Review Model (O'Hara et al., 2012). As noted in Boring et al. (2015), NUREG-0711 is written primarily for use by the U.S. Nuclear Regulatory Commission (NRC) in reviewing the HFE activities undertaken by the licensee. As such, the level of explanation provided in the guideline is not specifically calibrated to the needs of industry; rather, it serves as a quality check on the HFE process that industry should follow to achieve successful HSIs in the control room. This seeming disconnect between the purpose of the guideline and industry need for additional process guidance should not be seen as a deficiency of NUREG-0711. Additional guidance is readily available in supplemental reports commissioned by the U.S. NRC. For example, NUREG/CR-6393, *Integrated System Validation: Methodology and Review Criteria*, by O'Hara, Higgins, and Brown (1995) outlines the details pertinent to one of the major HFE processes that the U.S. NRC expects industry to follow. Additional guidance is also available from EPRI, e.g., EPRI 3002002770, *Guidance for Developing a Human Factors Engineering Program for an Operating Nuclear Power Plant* (2015), or even through HFE standards such as ISO 9241-210, *Human Centered Design for Interactive Systems* (International Standards Organisation, 2010).

The problem is not that there is a lack of relevant guidance on using HFE to support control room modernization. Instead, the problem is that there is actually an overabundance of HFE guidelines, methods, and processes from which to choose. This richness can in itself become another barrier to control room modernization by obfuscating the HFE process. Rather than becoming an enabling process, HFE risks becoming murky or overwhelming to the plant design engineer who is planning the upgrade. What is the best process for ensuring operator needs and wishes are met in the modernization process? What are the measures of operator performance in this process? What are the success criteria for HFE?

The goal of the LWRS Control Room Modernization Project is to demystify the HFE process and provide concise guidance to utilities and vendors to enable them to design and validate new control systems for the main control rooms of nuclear power plants. The present report focuses on one element of the HFE process, namely verification and validation (V&V) of the new system. V&V is the evaluation of the HSI according to operators or HFE standards. V&V is more than conducting an operator-in-the-loop study, and the process outlined here should adequately allow utilities and vendors to understand the nuances of V&V and adjust their process accordingly.

In February, 2015, the authors of this report were invited participants in the *Experts Workshop on Human Factors Validation of Nuclear Power Plant Control Room Designs and Modifications* hosted by the Office of Economic Cooperation and Development (OECD) Nuclear Energy Agency (NEA) Committee on the Safety of Nuclear Installations (CSNI) Working Group on Human and Organizational Factors (WGHOF). The OECD NEA CSNI WGHOF validation workshop report is forthcoming and will likely prove an indispensable addition to the literature that supports utilities and vendors in their V&V activities in support of control room modernization. One overriding theme that emerged in the workshop was the need for establishing reasonable confidence in the quality of the design and in the quality of the validation. Guidelines such as NUREG-0711 provide flexibility, but in doing so, they fail to provide a single process or conclusive acceptance criteria for validating designs. This report presents the authors' attempts to provide a thorough V&V process for HFE. It directly complements an earlier report, Operator Performance Metrics for Control Room Modernization:

A Practical Guide for Early Design Evaluation (Boring et al., 2015), which maps an HFE process and measures across NUREG-0711 requirements (see Table 1). The present report considers only V&V aspects of NUREG-0711. However, it should be noted that V&V activities as presented here span a wide portion of NUREG-0711 beyond the column dedicated to V&V.

Table 1. NUREG-0711 Process Model with Added Steps Appropriate to Control Room Modernization (from Boring et al., 2015).

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

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

The remainder of the report is organized as follows:

- Chapter 2 – presents the current evaluation process adopted by the nuclear industry and corresponding challenges and limitations

- Chapter 3 – presents the safety case as new perspective for the evaluation process

- Chapter 4 – presents GONUKE as a method for collecting evidence to build a safety case for licensing applications of control room upgrades

- Chapter 5 – presents future research for building a complete safety case of sufficient confidence to indicate safe plant operations

- Chapter 6 – summarizes the report.

(This page intentionally left blank)

2. VERIFICATION AND VALIDATION OF CONTROL ROOMS IN THE NUCLEAR INDUSTRY[1]

2.1 Introduction

V&V as applied in the nuclear power community has tended to focus on final evaluation in the licensing applications to the U.S. NRC. In the context of licensing review, NUREG-0711 (O'Hara et al., 2012) explicitly states that V&V "is considered a test that final design requirements are met" (p. 74), although HSI tests and evaluation are recommended during the design process. Consequently, both research and practice of V&V in the nuclear industry have emphasized final or summative evaluation of the HSI design. This late-stage evaluation is called integrated system validation (ISV) and is comparable to a factory acceptance test (FAT), except the acceptance criteria in ISV center on operator performance while using the system vs. software or hardware reliability in the FAT.

NUREG-0711 provides substantial details on the review of V&V in the licensing process (Chapter 11) in contrast to the short description of the HSI tests and evaluation (pp. 60-61). The NRC prescribes four key V&V activities:

(1) *Sampling of the Operating Conditions:* ensures that licensees identify the environment and potential situations that may arise during the actual operation of the plant, reflect system performance under those varying conditions, and examine significance of HSIs in those operating conditions. Effective sampling of operating conditions ensures that system safety inferred from subsequent V&V activities can generalize to the entire operating life of the plant.

(2) *Design Verification:* ensures that licensees design HSIs to support operators for the full range of operating conditions (i.e., sampled scenarios). This includes analytical evaluation of the HSIs using task analysis. Effective design verification analytically identifies HSI deficiencies or Human Engineering Discrepancies (HED) that must be addressed prior to integrated system validation or plant commissioning.

(3) *Integrated System Validation (ISV):* ensures that licensees validate the system performance necessary for safe plant operations over a range of operating conditions. ISV typically involves human-in-the-loop studies recruiting full-scope simulators and professional reactor operators to provide empirical, performance-based measurements. Effective ISV empirically identifies HSI deficiencies or HEDs that are missed in design verification and must be addressed prior to plant commissioning.

(4) *Human Engineering Discrepancy Resolution:* ensures that licensees resolve any design HSI deficiencies identified in the V&V process. Effective HED resolution ensures that the HEDs

[1] Portions of this chapter first appeared as separate white papers presented by authors at the OECD NEA CSNI WGHOF *Experts Workshop on Human Factors Validation of Nuclear Power Plant Control Room Designs and Modifications*. One paper is forthcoming (Boring, in press).

are eliminated prior to plant or system commissioning, eliminating risks of unsafe operations.

This chapter examines the limitations of adopting the conventional perspective that V&V should emphasize final evaluation, especially for the practitioners who must manage the engineering design and licensing process as well as the V&V process. The conventional perspective of V&V as ISV may be less suitable for the nuclear industry engaging in step-wise modernization projects that involve ongoing or gradual modifications to existing HSIs for plants in operations. Further, there are theoretical and practical limitations in obtaining best available empirical evidence for generalizing plant safety during operations from ISV results only. The first part of this chapter discusses the lack of early or formative evaluation, while the latter focus on the technical challenges of relying on ISV results for generalizing plant safety over the operating life cycle.

2.2 Challenges of Integrated System Validation

2.2.1 The Two-Edged Sword of Conclusiveness and Precision

Effective V&V in control room modernization must both meet regulatory rigor and maintain operational safety at the plant. There is thus a natural desire with V&V to achieve measurement precision and lawlike certainty of findings. Using the analogy of ISV as a type of FAT, there are several common types of FATs. For example, in the software industry, clear standards and acceptance criteria guide the process of V&V, from systematic debugging, to alpha and beta version testing of software by end users, to requirements-based testing (see ISO/IEC/IEEE 29119-1, 2013). Such testing is increasingly finding a place in sequential software development, whereby the software is tested at certain milestones in the development lifecycle. It is not clear why there remains such a heavy emphasis on ISV for control rooms.

To speculate, the persistence of favoring late-stage evaluation through ISV may result because practitioners and 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 is no room for error or refinement. The book is figuratively closed once the ISV is performed, and there is no need for further questions. Unfortunately, rarely are the findings from V&V so conclusive, even when they must stand up to regulatory scrutiny. As such, V&V practitioners and researchers must learn to accept some degree of uncertainty in the evaluation results. Humans are remarkably resilient to consistency and classification. There must be a degree of acknowledgement and even acceptance of the imprecision of V&V. Rather than rely on one conclusive ISV, a better approach might be 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 practitioners and especially researchers have tended to gather increasingly complex measures of performance. It might be argued that this is in pursuit of a more scientific and rigorous set of findings rather than the subjective measures most often employed in usability-type 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.

V&V researchers and practitioners strive to provide highly conclusive findings and try to do so through an ever increasing arsenal of measurement methods. These forms of V&V may not be productive or sustainable. Instead, the practice of V&V might be better served by embracing the evolving nature of the findings afforded by early-stage evaluation using reasonable measures to support the analysis. This report attempts to refine the processes and measures of V&V to best reflect operator performance and system interfaces. It is necessary to continue verifying and validating V&V.

2.2.2 Better Late Than Never?

It is accepted in the HFE community that it is better to be involved early in the design of a system rather than later (International Standards Organization, 2010). This stems from the best window in the design cycle for HFE 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 HFE as contributors to the end product, nor does it allow adequate time to fix issues that may surface in the system.

Boring et al. (2014) emphasize how existing guidance on evaluating HSIs in the nuclear industry has a strong emphasis on ISV, which is a late-stage evaluation on the completed design. It is natural that the regulatory review emphasizes summative evaluation. Because NUREG-0711 only minimally calls out early-stage evaluation, this may be interpreted by system designers to mean ISV is the only required or, indeed, preferred type of evaluation. Additional guidance provided by industry counterpart, Electric Power Research Institute (EPRI, 2005) carefully matches design and evaluation phases to NUREG-0711, thereby also emphasizing summative evaluation.

The nuclear community, with its strong emphasis on summative evaluation in the form of ISV, potentially puts itself in the position of doing HFE at the tail end of the design process, when HFE is, 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, as noted, is foremost a document guiding regulatory review at the completion of the design cycle rather than an exhaustive best practice for HFE. The propensity for late-stage V&V may also be a result of certain disclosure hesitancy between the licensee and the regulator, in which the intermediate steps of a design—the formative 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, design engineers have not necessarily engaged in a process of system improvement based on user input and evaluation. Nor have they documented lessons learned in the design process. Thus, HFE tends to focus on demonstrating that the overall system as designed actually worked. In this manner, however, HFE hasn't demonstrated that the design evolved to the point of working.

Design engineers may seek to rubber stamp the design through HFE rather than actively refine it.

In the opinion of the authors, it is necessary 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. The purpose of V&V is 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 HFE process.

In short, for HFE 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 HFE can help shape and optimize the design of the HSI leading up to ISV. ISV is the culmination of earlier HFE efforts, not a substitute for them. ISV confirms the design, but early-stage V&V holds the promise of improving the design.

2.2.3 Measures That Don't Measure Up

The use of the performance measures in V&V is sometimes driven by the state of the art in HFE, not by their practical utility. This statement must not be misinterpreted to be a criticism of the many solid HFE approaches and methods 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 the measures actually measuring what they 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. The field of HFE needs 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, HFE risks 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 actually needed for operator performance. Sometimes straightforward usability measures (Tullis & Albert, 2008) may serve the needs of HSI evaluation adequately.

Again, the purpose here is *not* to criticize research that uses these types of measures, which may in fact be the key to understanding operator performance better. These and any number of advances in psychological measurement do not necessarily help perform V&V better than is currently the case. The V&V practitioner and researcher must stop and determine how different measurement tools available actually help clarify and model operator performance. If the measures do not specifically verify or validate, they should be discarded or refined. V&V activities must not be distracted by a gluttony of measurement options.

It is useful to co-opt the term ALARA,[2] here meaning *As Low As Reasonable Assessment*. The field of usability engineering is instructive for framing the ALARA concept. While extensive and elaborate methods for assessing the usability of a system abound, there is also a movement toward discount usability (Nielsen, 1995). Discount usability emphasizes a collection of methods that may be applied whenever feasible. In other words, some form of user evaluation is better than foregoing evaluation because of inadequate time and resources to perform the ideal evaluation. Even informal evaluations early in the design stage can help shape the design positively. Discount usability methods are now being integrated in agile software development as a responsive, inexpensive means to gain design feedback (Kane, 2003). Unfortunately, there is currently no ready place in the conventional V&V framework to capture such quick and informal design evaluations in support of control room modernization. A graded approach to evaluation is desirable, and the benefit of such evaluation should be credited toward meeting licensing requirements. Even safety-critical systems like those found in the control room may benefit from iterative feedback leading up to a formal ISV in terms of improving the control room design and developing the summative evaluation. What is missing is guidance of the acceptability of adopting ALARA in a heavily regulated industry like nuclear power.

2.2.4 Leaving Everything to the End

As noted, the desire for a definitive conclusion on system performance for licensing has led to tremendous focus on ISV. ISV is designed to yield the performance-based measurements for indicating sufficient safety over the permitted operating period. In the nuclear industry, ISV typically employs full-scope simulators and professional operators to represent the final control room/systems design (i.e., hardware, software, procedures and personnel elements) for conducting a series of performance-based tests. In theory, this approach can produce the necessary (or at least most ecologically valid) empirical evidence of system performance that should be generalizable over the full licensing period.

In practice, plenty of challenges are associated with ISV that employs full-scope simulators and professional operators to collect sufficient evidence of system safety for generalizing over the entire licensing period. ISV has many practical constraints in testing integrated operations of the control room that lead to many technical challenges in collecting and analyzing sufficient data to conclude system performance over the licensing period. The key, well-known practical constraint of performance testing is the labor requirement on two types of personnel—operator crews and experimental staff. For integrated testing of control room design, operator crews represent a key element, but they are always in demand for other activities. Consequently, the number of crews and time of each crew allocated for ISV are limited, or at least kept to the minimum. The availability of operator crews is unlikely going to improve, posing continual challenges in generating sufficient performance data for drawing performance conclusions. The shortness of crews poses a number of problems for ISV:

[2] ALARA has historically meant As Low As Reasonably Achievable, referring alternately to minimizing accidents or radioactive exposure. Another variant is As Low As Reasonably Practicable (ALARP).

- *Sample size.* The availability of operator crews for ISV poses three kinds of technical challenges for analyzing test data and thus drawing safety conclusions. The first two challenges concern conclusion validity due to inadequate sampling (for both qualitative and quantitative methods). A limited sample size or number of crews participating in ISV activities may inadequately account for intra-crew variations, which is analogical to individual differences. Further, from a statistical perspective, crews should be randomly selected for appropriate generalization (i.e., random effect models) but this criterion is rarely satisfied strictly. Subjective and "objective" (i.e., statistical) correction could be applied to qualitative and quantitative data to moderate impact. However, the validity of correction methods is never examined for performance based testing in nuclear process control. In brief, inadequate sampling of operator crew could impact the validity and confidence in drawing performance conclusions on the integrated control room design.

- *Scenario sampling.* The second challenge concerning validity of ISV results is inadequate sampling of scenarios. Operator crews need to participate in a significant range of scenarios in order to provide the range of data to provide performance estimates or conclusion for all operating conditions. Further, from a statistical perspective, scenario selection, like operator crew recruitment, should be random for generalization, but this criterion can be difficult to satisfy strictly. The limited availability of operator crews for performance testing implies that the number of sampled scenarios is also limited. Consequently, the generalization of test performance to operating performance has limitations, irrespective of quantitative and qualitative evaluation methods when operator availability is limited.

- *Statistical power.* The final challenge associated with limited operator crew availability and thus performance data is the lack of statistical power for quantitative methods. Full-scope simulator studies typically require over 10 data points per experimental condition to indicate significant difference of medium-size effects. Additional data points are typically required for equivalence testing (that may be used in benchmarking studies). In brief, limited availability of operators constrains the traditional applications of inferential statistics on ISV studies for drawing performance conclusions.

The experimental (or performance testing) staff represents another set of practical constraints. The typical experimental staff involves human factors experts, simulation engineers, operations experts (who may be instructors or operators), and process experts (who may be plant engineers or vendors). In particular, ample interaction time between human factors professionals and process experts (i.e., experienced operators) are critical to develop effective scenarios and performance criteria. The scenarios required for V&V differ from the scenarios typically employed in operator training, and new scenarios are likely necessary to test operator interactions with the system rather than teach operating concepts. Further, performance measurements often involve some expert ratings. However, the availability of this interaction appears to be typically constrained or underestimated as process experts often have other duties and limited exposures to running performance testing from an ISV and experimental perspective. Most HFE professionals do not have frequent opportunities in full-scope simulator evaluation, while most experts are experienced with testing mainly from a training and examination perspective. In brief, the interaction time between HFE professionals and process experts can have a major impact on quality of testing scenarios and measurements. The necessary mutual learning curve for both parties is not always fully appreciated in the V&V budget or schedule.

The dependence on the interactions between HFE professionals and process experts for quality ISV performance measurements has three technical implications.

- First, interactions between HFE professionals and process experts can drastically improve sensitivity and reliability of the performance measurements that provide the necessary statistical power for drawing conclusions (Lau et al., 2014).

- Second, some expert judgments, and thus potential biases, are often embedded into human performance data. The details of these expert judgments or biases in the human performance data are often unknown to the researchers or data analysts. In addition, the quality of expert judgments is a function of multiple factors, such as the type of scenario (e.g., design basis event vs. beyond design basis event) and personal preferences (e.g., risk aversions). Consequently, measurement errors are not necessarily constant, let alone individual differences or reliability between experts.

- Finally, the level of necessary interaction between HFE professionals and process experts may limit degree of independence practically achievable between the evaluation and engineering teams in modernization projects. That is, the engineering team who has intimate knowledge of the control room technology may be highly effective at developing scenarios and measures for assessing the design but are recommended not to be integral part of the evaluation due to potential bias. While, there is the potential for the engineering team to bias the outcome of the evaluation, the complexity of the control room and of the specific systems being upgraded practically requires the help of process experts to be a part of the evaluation team to develop an effective control room assessment..

In brief, confidence in safety conclusions from ISV study is actually a complex, multivariate construct.

In nuclear process control, ISV also has two interrelated technical constraints that pose challenges in analyzing data and drawing safety conclusions. The first technical constraint is the combinatorial explosion of scenarios (or even scenario types) due to plant complexity. Formal methods do not exist to determine how interaction of components leads to qualitatively different scenario types nor what portion of all possible scenarios are covered by a particular set of scenarios. Thus, at least formally, the content validity or comprehensiveness of performance testing is difficult to assess. The second technical constraint is simulator fidelity of process behaviours for beyond design basis or severe accident events such as occurred in the seismic and tsunami flooding event at Fukushima Daiichi. The unknowns in the behaviors of the nuclear process during rare events prompt validity and reliability questions on all human performance measurements. The current resolution to these two technical constraints is largely based on the practical and subjective recommendation of experts. Adopting the practical stance of relying on process experts is probably necessary to maintain plant operations until researchers or practitioners can develop substantially improved solutions to problems of scenario sampling and unknowns in beyond design basis accident scenarios. This practical approach needs to account for the variability associated with process expert judgment for drawing safety conclusions from

ISV results. However, variability of process expert judgment is rarely examined carefully in the nuclear industry.

To summarize, the dependence on experts to develop complete and representative scenarios and the limited availability of operator crews to generate performance data limit the feasibility, or at least the validity, of direct adoption of well-established scientific techniques to determine plant safety over the licensing period. The limitation in performance data is compounded by the complexity of nuclear power plants that practically have infinite numbers of operating scenarios. The classical performance measures and analysis techniques advocated in the ISV literature often do not provide adequate strength of the evidence or confidence in plant safety assessment.

2.2.5 Evaluation Theory and Practice Are Not the Same

The majority of data collection and analysis methods in science are developed (initially) with considerations of neither the industrial purpose (and safety implications) of ISV nor inherent constraints of the nuclear domain. Scientific research methods for performance assessment originating in psychology and physiology rely on a large participant sample size to investigate many narrow research questions (cf., scenario types) and thereby produce knowledge or generate discussion for further testing and validation. In addition, methodological limitations often become impetus for further research. For instance, qualitative methods can focus on single participants to explore details and contexts with limited emphasis on generalization. Quantitative methods can focus on strict statistical or other criteria with large samples for validation. From this perspective, science is a continual process, readily accommodating *half-answers* to a research topic in anticipation of future studies.

The nuclear industry cannot accommodate *half-answers* to ISV or safety assessment. For instance, regulators cannot grant licenses on the basis of perfectly validated safety performance for only half the operating conditions. However, science classically produces research methods and study designs that produce conclusive or highly confident narrow findings (i.e., *half-answers*). This approach is impractical for ISV given the constraints of the nuclear industry. Consequently, both researchers and practitioners raise questions on classic research methods for ISV. Within the topic of ISV, the methodological discussions range from meaningfulness of inferential statistics and relevance of qualitative measurements to requirements for follow-up evaluation studies for validation of the main control room. The nuclear industry must address the applied research issues of ISV, since the direct application of classic research methods may not practically provide the necessary evidence and thus confidence in the assessment of plant safety.

2.3 Revisiting Verification and Validation

The various issues raised in this chapter speak to the challenges of performing a common V&V process or performance testing of integrated control room operations in nuclear power plants. It is difficult to acquire sufficient evidence and reasonable confidence in the results for plant safety assessment. Confidence in the late-stage V&V process, analysis, and results has major implication in the engineering design process as well as regulatory licensing decisions. Interestingly, confidence in performance testing connects closely with the longstanding basic research on test validity and validity generalization. Validity research provides a new perspective to revisit the purpose of V&V in the nuclear industry that may simplify the discussion and provide directions for future research.

Research questions on test and assessment validity have evolved over time (Murphy, 2009). Prior to the 1970s research focused on *Which forms of validity should be used?* In the recent past between 1980 and 2000, research resolved the issue of *Is it valid?* with meta-analyses. Presently, research is looking into *Validity for what purpose?* from a multivariate perspective. A solution may be valid for one purpose but not another, and validity is often a tradeoff to meet particular constraints. The evolution of validity research, particularly the present phase, puts ISV work into perspective. Prior to formulating any research programs to improve confidence in the findings from ISV, the nuclear community must answer *Drawing conclusions for what purpose?* The general answer to the question should be simple, though detailed versions may be contentious. Test validity is a multivariate construct; thus, one central statement, no matter the level of emphasis, is insufficient to represent the full concept for ISV in the nuclear domain. One pertinent area missing in the statement is that the resource requirements must be balanced with confidence in the predictions. If resource requirements are completely ignored for perfect predictability, then no nuclear power plants would ever be licensed nor would any control room upgrade ever be approved. Such a consequence invalidates the target outcome of performance testing.

To resolve this quandary, the following statement is proposed for drawing conclusions in V&V performance testing:

> *Conclusions from performance testing of integrated control room operations should help decide whether the integrated control room design can support reasonably safe plant operations over the requested licensing period (e.g., 20 years extension).*

While the proposed answer is no epiphany to anyone familiar with V&V of main control rooms, the statement focuses effort and results of V&V toward the goal of predicting safety over a number of years. This focus can help simplify or reframe the questions and discussions. Assuming agreement with the above statement, V&V is concerned with predictive validity. Thus, broadly speaking, all qualitative or quantitative data are collected and analyzed to make (or to become confident in making) inferences on future performance. Though the traditional assessment criteria may be impractical, many principles in inferential statistics (e.g., Type I and II errors[3]) remain essential for establishing confidence in V&V performance conclusions. More importantly, this perspective encourages all evidence that could support the prediction to be admitted for plant safety assessment, even though individual pieces of evidence carry different merits that must be carefully weighted for achieving valid conclusion on the integrated operational performance. This approach of emphasizing predictive evidence stands in contrast to treating the V&V activity and results solely based on testing the final design for a highly defined set of human performance metrics. Instead, V&V should steer toward establishing confidence in plant safety over the licensing period. This approach suggests the value of many types of

[3] A Type I error is a false positive, and Type II error is a false negative in statistical hypothesis testing.

evidence to establish the trajectory rather than a single all-inclusive snapshot of performance through ISV.

3. THE SAFETY CASE FOR VERIFICATION AND VALIDATION

3.1 Introduction

The paucity of early design evaluation and the limitations of final performance testing motivate an investigation into new approaches and methods for V&V in the U.S. nuclear industry, especially for those plants undergoing modernization. In particular, it is important that any new approach to V&V must encapsulate the perspective of *consequential validity*.[4] Consequential validity emphasizes the implications of the decision made as a result of the outcome of the evaluation method (i.e., V&V). For the practitioners of the nuclear industry (e.g., vendors, utilities, and regulators) who focus on outcome of a specific instance of V&V, consequential validity may be viewed as predictive validity—validity of the empirical evidence for predicting safety for the licensing period of the specific plant. For V&V researchers developing generalizable evaluation methods relevant to multiple modernization and construction projects, consequential validity extends beyond predictive validity in that the developed V&V method can have major decision implications on licensing, plant operations and ultimately public safety. The focus on consequential validity ensures an emphasis on evidence predictive of safe plant operations over evidence centered on performance testing a high-fidelity representation. This perspective reconciles the need for early evaluation and importance of final integrated testing by weighing the relative merits of evaluation at different stages and the quality for predicting safety in the future. That is, empirical evidence at early or formative evaluation *likely* qualifies less than final or summative evaluation for predicting safety in the future. Nevertheless, evidence from formative evaluation may be sufficient to predict many aspects of safety or system performance, including establishing the trend toward improved performance. Thus, this chapter presents a *safety-predictive* approach to V&V that accommodates evidences at different stages of

[4] Messick (1995) articulated that consequential validity, in the context of education, "appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice." In consequential validity, the evidence and activity concerning V&V should be built around the consequence of the decision being made, which can have implications beyond the V&V exercise. For example, validating the selection of a particular input device as part of a control room modernization activity (e.g., Ulrich, Boring, and Lew, 2015) can have implications beyond the simple system for which it is deployed. The results of the V&V may be used to justify further implementations for other systems at the plant or even, if results are shared across the industry, to shape the preferred industry input device. Predictive validity does not conceptually focus on the decision consequences of the evaluation effort. For instance, overemphasis on predictive validity in V&V method (i.e., perfect measurements) may lead to overly stringent licensing criteria, and thus result in safe-enough plants being denied operating licenses. Such a V&V method violates consequential validity, as safe-enough plants are denied operations to produce electricity and thus increase costs for the public. V&V practitioners and researchers should consider consequential validity, because the method and findings can have consequences on the entire industry and the public.

evaluation. This emphasis on safety prediction should improve the licensing process for the utilities and the confidence in the licensing decisions for the regulators.

To support the nuclear industry in adopting a safety-predictive approach that leaves behind the notion of V&V concerning only the final design, four methodological areas require substantial development:

1. Structuring the evidence, especially the information gathered outside of the traditional V&V process

2. Identifying the appropriate evidence to gather with respect to different stages of design and evaluation

3. Assessing the merits of various evidence for predicting plant safety

4. Integrating all of the evidence to provide a final safety assessment of integrated operations in the main control room.

Clear and effective methods in these four areas can ensure that the process of gathering, assessing, and presenting evidence would lead to products that could satisfy regulatory concerns on public safety.

3.2 The Safety Case

This chapter presents the safety case as a meaningful starting point for structuring evidence to encapsulate the concept of consequential validity. A *safety case* is defined by Kelly as follows (1998, p. 22): “A safety case should communicate a clear, comprehensive, and defensible argument that a system is acceptably safe to operate in a particular context.” The three elements of a safety case as described by Kelly and depicted in Figure 1 are:

- *Requirements*—the safety objectives for the system

- *Argument*—the mapping of the evidence to the requirements

- *Evidence*—supporting safety documentation such as risk assessments

The exact nature of the evidence and the way in which arguments are most effectively conveyed is the subject of ongoing discussion (e.g., Haddon-Cave, 2009). Nonetheless, they remain the cornerstone of safety regulation, for example, in the United Kingdom (UK) defense (UK Ministry of Defence, 2015) and nuclear power sectors (UK Office for Nuclear Regulation, 2013). Rather than rely on a single source of evidence for the safety of a system, safety case regulation requires a body of evidence that clearly argues for meeting safety requirements. This approach is not unlike the judicial trial system, where evidence must be argued to influence the verdict. The verdict, in this case, is the safety of the system.

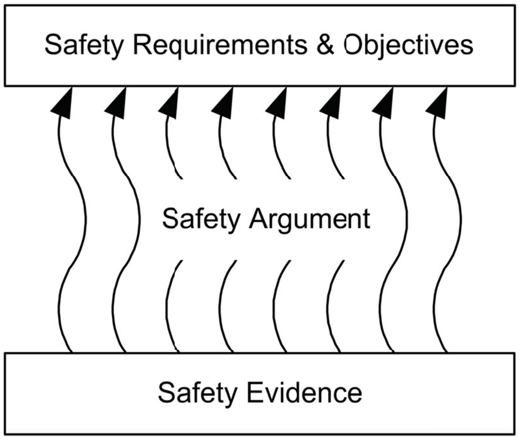

Figure 1. Composition of the Safety Case (from Kelly, 1998).

Kelly (2008, p. 31) notes, "Safety cases have little hope of adding value if they are impotent in their influence on the design and operation of the system in question. Safety cases shouldn't be produced after the system design has been finalized." Thus, the argument can be made that safety cases should not rely on late-stage evidence but rather on evidence derived early in the design stage and actually used to shape the design of the system. This approach seems to contradict the late-stage emphasis of ISV. That is not to say that ISV is an unimportant piece of evidence in the case for safety; however, it isn't the only evidence used.

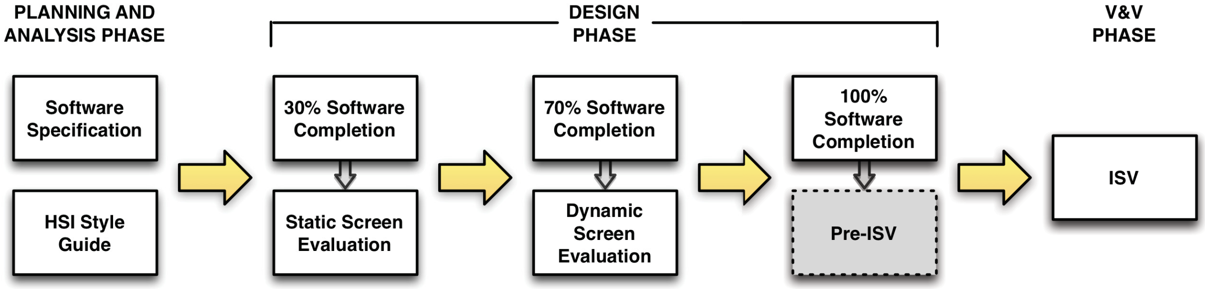

Figure 2. Design Phase Evaluation (from Boring et al., 2015).

Figure 2 provides one example of evidence in the form of preliminary usability evaluations during the design phase. Recall that NUREG-0711 emphasizes the primary form of evaluation as a standalone ISV in the V&V phase. In the figure, currently being used to support control room modernization activities at a fleet of U.S. nuclear plants, there are actually three rounds of assessment that occur prior to the ISV. These consist of operator-in-the-loop studies and expert evaluation by HFE professionals. The three phases correspond to the 30%, 70%, and 100% system completion milestones, resulting in different fidelities of the system being tested. However, each phase of evaluation serves as input for the next design and development activities for the system. As such, the system undergoes an iterative design-evaluation cycle leading up to the completion of the system.

The process outlined in Figure 2 illustrates a systematic design process by the licensee and vendor, one that takes operator and HFE input at several junctures during the design and uses them to refine and optimize the design. This process might even be said to exemplify a user-centered design best practice. Yet, these iterative pieces of evidence do not have a clear place in the NUREG-0711 framework. They are supplemental evaluations leading up to the ISV, which is the truly meaningful evaluation in the common interpretation of NUREG-0711.

What if these design-phase evaluations were not just supplemental steps in the process? What if they were framed as a type of safety case—evidences that build the argument for meeting safety objectives for the design of the new system? Surely there is value in the design-phase evaluations beyond guaranteeing the system will pass the ISV. The design-phase evaluations are more than a dry-run for the final evaluation, because they are actually shaping the design of the system. These evaluations become evidence for the veracity of design decisions and the quality of the process of the design. In other words, they provide data to show why one design decision was selected over another but also build confidence that the final design represented the convergence of a vetted process. By using iterative data, it becomes clear that passing the ISV is not a matter of luck; it follows a traceable path since the design inception. Unfortunately, when the regulatory process emphasizes evidence coming from the ISV, there is no clear guidance for the licensee and vendor to build a complete safety case with evidence across the design lifecycle.

If the case has been made for the value of early-stage evaluation, is there still a need for late-stage ISV? There are several key differences that *may* help delineate early vs. late stage evidence:

- Scenarios—the situations against which the system is tested

- Participants—the representative sample of operators who will interact with the system

- Measures—the reflection of operator performance and preferences using the system.

Table 2 summarizes these considerations. Essentially, the distinction between early and late stage evaluation can be understood in terms of completeness and conclusiveness. Late-stage evaluation, as a type of final evaluation before licensing, will seek to have complete scenarios against which the system is tested, a representative and statistically valid sample of operators, and rich, definitive performance measures. Early-stage evaluation may feature a subset of these tailored toward gathering sufficient data for meeting design objectives. The goal of early-stage evaluation is ALARA, while the goal of late-stage evaluation is sufficient assessment to meet certification for regulatory safety requirements. The fact that early-stage evaluation may not be a rigorous as late-stage evaluation does not diminish its value as evidence toward completing a safe final design.

Table 2. Considerations of Early and Late Stage Evaluations.

		Evaluation Stage	
		Early Stage	**Late Stage**
Evaluation Considerations	**Scenarios**	Generally limited to test the functionality and operator interaction with specific aspects or features of the system	Complete across the range the system will encounter relative to real world situations, including unlikely but safety-critical worst case situations
	Participants	Limited number of operators and process control experts needed to test the evolving system design and provide feedback to the design team	Ideally, a large enough sample size to be statistically significant, covering a range of operators (e.g., different experience levels) representative of the operator population
	Measures	Suitable for design decisions, akin to discount usability, with consideration of subjective preference data to drive the design	Suitable for safety compliance decisions with emphasis on objective measures of performance

These considerations apply only to operator-in-the-loop studies. The next chapter introduces other forms of evaluation that may prove more qualitative in nature than empirical studies. The need for clear appreciation of different types of evidence becomes more importunate when the evidence is not strictly numeric. Other forms of evaluation may not produce quantitative assessments. The safety case argument is crucial for incorporating qualitative evidence that may result from the V&V.

Despite its prevalence in the European regulatory community and others, the safety case has not been widely adopted in the U.S. Recently, there has been consideration of safety cases to help regulate the U.S. chemical industry by the U.S. Chemical Safety Board (Hopkins, 2013), and the U.S. Food and Drug Administration has introduced safety assurance cases (which are in most cases synonymous with safety cases) to minimize risk in the use of medical devices (U.S. Food and Drug Administration, 2010). These industries are considering safety cases as part of a risk regulatory framework. In the U.S nuclear industry, the risk regulatory framework is guided by probabilistic risk assessment (PRA), including human reliability analysis (HRA). HRA has diverged from HFE in the U.S. (Boring and Bye, 2008), and the two approaches may not initially seem compatible. Even the most recent revision of NUREG-0711 (O'Hara et al., 2012) removes the explicit connection between HRA and the HFE process, whereas the preceding version of

NUREG-0711 (O'Hara et al., 2004) coupled the two disciplines. If the HSI safety case is most closely associated with HFE, there remains a potential separation of safety and risk that may need to be resolved before the safety case serves as more than a supplement to regulatory requirements.

It is important to note that although NUREG-0711 emphasizes ISV, it does not prohibit other stages of evaluation. Licensees and vendors will need to provide examples of such evaluations in order for the regulator to determine their effectiveness. Such examples must not be fragments that confuse the merits of the safety case. They should proceed in a systematic manner that builds an effective argument. The next chapter introduces an approach to gathering different types of evidence across the system lifecycle. This approach can help ensure that evidence is reasonable and comprehensive. Although the U.S. nuclear regulatory framework does not currently require safety cases, this framework certainly does not discount the value of well-argued safety evidence.

4. GUIDELINE FOR OPERATIONAL NUCLEAR USABILITY AND KNOWLEDGE ELICITATION (GONUKE)[5]

4.1 Introduction

The U.S. DOE LWRS Program Control Room Modernization Project has been working with commercial nuclear utilities to ensure that new digital systems installed in a control room are optimized to maximize operator performance (Boring et al., 2014; Boring & Joe, 2014; Hugo et al., 2013; Ulrich et al., 2014). As new digital replacement technologies have been introduced into control rooms, they may represent significant variability in the HSI due to differences in the digital systems deployed (e.g., generational differences in digital systems or general stylistic differences between different vendors) or even due to differences in the developers of the systems (e.g., inconsistency of implementation of HSIs within the same digital platform). By developing a consistent HFE process to be used across multiple system upgrades, it is anticipated that utilities will be able to standardize the digital HSIs as they are introduced in a gradual, stepwise fashion (Boring & Joe, 2014). This chapter highlights the process developed for control room modernization at INL. Herein, we label this process the *Guideline for Operational Nuclear Usability and Knowledge Elicitation* (GONUKE). While GONUKE is intended primarily for nuclear power plant control room modernization, the process may easily be generalized to other safety critical system applications.

4.2 The GONUKE Process

Individual aspects of the GONUKE process have been introduced previously in (Boring et al., 2014 and 2012). The reader is referred to those papers for more in-depth discussions of the process elements. This section highlights the key aspects of the method and describes how they are applied as part of a cohesive approach.

A key concept of the GONUKE process is the articulation of types and phases of evaluation. The purpose of *evaluation* is to demonstrate that the users (i.e., operators) of the system are able to perform tasks successfully. This may suggest only empirical evidence (e.g., a control room simulator study) is informative to evaluation. In fact, there are three types of evaluation that are helpful to establish the success of a design:

- *Expert Review:* This is *verification*—evaluation of the system by subject matter experts against a standard set of criteria.
- *User Testing:* This is *validation*—evaluation by testing operator performance in actual use of the system.
- *Knowledge Elicitation:* This is capturing the epistemic or knowledge insights of the operators who use the system—what we will here coin *epistemiation* (pronounced: ˌɛpɪˌstiːmɪˈeɪʃən).

[5] This chapter is based on a paper by Boring, Ulrich, Joe, and Lew (in press) presented at *the 6th International Conference on Applied Human Factors and Ergonomics and the Affiliated Conferences (AHFE 2015)*.

Epistemiation results from the operators providing their experience and recommendations while using the old and new systems. While verification entails evaluation of the system against established human factors standards by HFE experts, epistemiation centers on the expert users (i.e., the reactor operators) and their hands-on knowledge of how the system can and cannot be used. The experience of the users of the system also goes beyond validation, which focuses primarily on measurable aspects of operator performance. Epistemiation consists of the qualitative insights by actual users who articulate, based on their subjective experience, the disparities between the system as it is and the system as it ought to be. These insights by operators are not typically framed in terms of concrete design recommendations or grounded in HFE principles. They are nonetheless invaluable in shaping the system. Epistemiation in the sense of gathering expert user feedback is not one of the common tools of usability engineering and user-centered design (Rubin, 1994), and its potential application in supporting evaluations is still being explored and developed at this time.

The evaluation *phase* refers to when the system is evaluated. As noted throughout this report, the regulatory framework (i.e., NUREG-0711) tends to emphasize evaluation of the completed design. This late-stage evaluation in valuable, but it has several disadvantages:

- To be conclusive in verification, it may require an exhaustive review against human factors standards with literally thousands of relevant review criteria.

- To be conclusive in validation, it may require a large number of participants to have sufficient power to be statistically significant (see Section 2.2.4 for discussion).

- There is no room for error on behalf of the operators and the system, as human factors issues identified on the final design may delay deployment of the system and require costly reworks.

There is merit in performing evaluations at earlier phases of the design. The phases of evaluation can be thought of as formative and summative, a concept borrowed from the field of education, where it has been used extensively to catalog teaching vs. program effectiveness (Scriven, 1967). Within human factors, these two phases of evaluation have been defined as (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.

Formative evaluation overcomes the earlier noted limitations of late-stage (i.e., summative) evaluation. It works to help refine the design before it is finalized, thereby ensuring a successful outcome at the summative evaluation phase. It also helps to build a safety case (Office of Nuclear Regulation, 2013) for the design of the system (as detailed in Chapter 3), providing evidence that the design works successfully. Formative evaluation establishes the trajectory of the design in terms of meeting HFE objectives of the system. As successive revisions of the system are evaluated, there should be a tendency to see fewer human factors issues such as operator errors and greater operator satisfaction with the use of the system. These iterative design-evaluation cycles at the formative stage establish that the system is improving throughout the design cycle and arriving at a safe, efficient, and usable system by the time it reaches summative evaluation.

Evaluation does not necessarily commence at the formative phase and end at the summative phase. There is considerable preliminary evaluation that goes into the planning and analysis prior to the system design. Likewise, after the system is implemented, there is ongoing monitoring of the system and operator performance. These phases correspond to what here is called *Pre-Formative* and *Post-Summative* phases of evaluation, respectively. In practice, Pre-Formative and Post-Summative might be considered outside the scope of evaluation, but they represent part of a continuum of evaluation that should be on-going rather than confined to discrete phases surrounding design activities. Especially Post-Summative evaluation should continue throughout the system lifecycle, whereas the other phases of evaluation are prompted only by system changes that require new design or design modification efforts.

Table 3. Phases and Types of Evaluation in the GONUKE Process.

		Evaluation Phase			
		Pre-Formative ***(Planning and Analysis[1])***	**Formative** ***(Design[1])***	**Summative** ***(Verification and Validation[1])***	**Post-Summative** ***(Implementation and Operation[1])***
Evaluation Type	**Expert Review** ***(Verification)***	**[1]** Design Requirements Review	**[2]** Heuristic Evaluation	**[3]** System Verification	**[4]** Requalification against New Standards
	User Study ***(Validation)***	**[5]** Baseline Evaluation	**[6]** Usability Testing	**[7]** Integrated System Validation	**[8]** Operator Training
	Knowledge Elicitation ***(Epistemiation)***	**[9]** Cognitive Walkthrough (Task Analysis)	**[10]** Operator Feedback on Design	**[11]** Operator Feedback on Performance	**[12]** Operator Experience Reviews

[1] Corresponding Phases in NUREG-0711.

The types and phases of evaluation are summarized in Table 3 and described below. Note that the phases of evaluation align with the four phases of NUREG-0711 (O'Hara et al., 2012) (i.e., Planning and Analysis, Design, Verification and Validation, and Implementation and Operation). The four phases and three types of evaluation of the GONUKE process comprise 12 possible steps of evaluation:

1. *Pre-Formative Verification:* Completed prior to the design phase by expert review. At this phase, the verification consists of expert input into the planning and analysis of the design. The human factors expert may review design requirements and provide preliminary design

recommendations. The human factors expert may also formulate an HSI style guide to shape the subsequent design phase activities.

2. *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 heuristic usability checklist (Ulrich & Boring, 2013; Boring et al., 2006).

3. *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.

4. *Post-Summative Verification:* Completed after deployment by expert review. This activity involves ongoing maintenance of the system to applicable standards. Human factors standards continue to evolve over time as knowledge about HSIs is refined and as new HMI technologies are invented. While the system may remain essentially unchanged over long durations, it is advisable to be aware of the implications of changes in the standards. Even where the system is grandfathered to an earlier standard, any future change to the system will likely ultimately require conformance to current standards. A periodic review of changes to standards and identification of gaps between the system and those standards can ensure that the system remains compliant and that upgrades and updates are unencumbered by a standards compliance barrier.

5. *Pre-Formative Validation:* Completed prior to the design phase by user testing. At this phase, a baseline evaluation should be completed. 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. The baseline evaluation provides the basis for benchmarking the new system against the existing system.

6. *Formative Validation:* Completed during the design phase by user testing. Typical for this type of evaluation would be usability testing of a prototype HSI (International Standards Organization, 2010). Formative validation is not typically a single evaluation (e.g., a single control room simulator study) but rather a series of evaluations performed in an iterative manner throughout the design phase of the system. The design in this manner is systematically improved as it approaches final implementation.

7. *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) and elsewhere (O'Hara et al., 1995). Integrated system validation is akin to the factory acceptance test for operator performance. The finalized system is tested using operators prior to deployment.

8. *Post-Summative Validation:* Completed after deployment by user testing. User testing may prove a bit of misnomer for the most frequent form of Post-Summative Validation, which is operator training. Training typically occurs one in six weeks on shift for licensed operators. While different systems are trained in rotation, an important component of training is interactive testing of operator performance and feedback from instructors to the operators on their performance. Performance during training is trended over time, e.g., (Chang et al., 2014), making it possible to have periodic validation of the operators' interaction with the system. Performance issues are noted for corrective action, either through additional training or, rarely when warranted, through changes to the underlying system.

9. *Pre-Formative Epistemiation:* Completed prior to the design phase through operator feedback. Operator knowledge can be elicited through formal queries to help capture the tasks and associated requirements of performing specific system activities. Such an analysis would be typical of a cognitive walkthrough used in support of a task analysis (Diaper & Stanton, 2004). These inputs would serve as design inputs to capture operator needs and expectations from the system.

10. *Formative Epistemiation:* Completed during the design phase through operator feedback. In tandem with usability testing (either as a piggybacked or as a standalone evaluation activity), operators can be polled on their experience with performing activities using the new system designs. Feedback can, for example, consist of explanations of operator expectations for data displays and particular indicators at certain steps in a process.

11. *Summative Epistemiation:* Completed after the design phase through operator feedback. While Formative Epistemiation calls for feedback on the design of the system, Summative Epistemiation elicits feedback on the performance of the system and self-assessment of each operator's own performance. This feedback can fine-tune any remaining design issues or identify human engineering deficiencies that have endured or emerged past the design phase.

12. *Post-Summative Epistemiation:* Completed after deployment through operator feedback. This can be facilitated through operator experience reviews. Note that this is different from *operating* experience reviews, which look at system performance. Operator experience reviews are periodic assessments of the operators' experiences using the system with a specific goal to identify areas

4.3 Special Considerations

An important element of epistemiation is the separation of user wants vs. user needs (Lindgaard et al., 2005). Epistemiation is not a focus group activity with the goal of gathering a design wish list from operators. Rather, it is a systematic attempt to capture operator knowledge throughout the system lifecycle. Epistemiation goes beyond user needs assessment, which is often centered exclusively on the design of a new system rather than enhancing or modernizing an existing system. In control room modernization, the operators are the true experts of the system processes, whatever the implementation, whether an existing system or a proposed new system.

The goal of epistemiation is to ensure that the system implementation matches the operators' mental models of how the system should work. In some cases, operators may propose additional features to the system. These wishes should be carefully balanced with the practical constraints of how the system is implemented (e.g., not all automation functions are practicable or acceptable from a regulatory perspective). In other cases, operators may not know what new features are possible, and epistemiation facilitates discussion between the operators and the system engineers building the new system. Expert users of a system are ultimately the individuals most qualified to provide design inputs, and epistemiation attempts to ensure there is a mechanism to include their design ideas beyond what would emerge from expert reviews or usability testing.

User testing or validation should not be thought of solely as a control room simulator study (Boring et al., 2015). Having operators execute scenarios to test the system and their response to the system is an effective way to collect insights on the system. In the context of modernization, such studies often comprise a benchmark comparison between the existing system and the modernized system (Boring & Joe, 2015). Benchmark studies use standard usability measures—such as time and accuracy to complete the task—as well as operator preference, to establish the efficacy of the new system either by matching or exceeding the operator performance of the existing system. However, the data obtained from a study should not be limited to those pre-scripted measures that produce numeric results. An important component of operator studies is the expertise that operators bring in using the existing systems. Open-ended feedback from semi-structured interviews will elicit operator knowledge that may be used to refine the system design to match the operator's mental models of the system. Legacy conceptions of how a system ought to function should not serve as limiters on the design of improved functionality and interface quality. Where the old way of using a system seemingly interferes with the new way of using the system, it is imperative that the HFE professional determine the merit of existing approaches and the potential for carrying those forward into the new design. When there are clear clashes between existing and new operational approaches and when the new approach represents advantages in terms of usability or safety, these design catch points become the basis for establishing training to override previously learned use biases.

There may be some hesitancy on behalf of the utility to release the results of the summative evaluations as part of a license submittal to the regulator (Boring, 2015). Where there is an emphasis on demonstrating the successful implementation of the system, it may seem counterintuitive to include information from the formative stages, which may include evidence that the system or operators were not successful. A shift in approach is critical for both the utility and the regulator to see the design process—including inevitable early-stage problems with the HSI—as evidence of a comprehensive process. HFE issues that were identified early and corrected prior to implementation of the system do not represent shortcomings or weaknesses in the design. Rather, the fact that issues were identified and corrected suggests an effective human factors process, which demonstrates the system is converging upon an optimal solution for the operational context.

Not all evaluations require all types and phases as depicted in Table 3. Certainly, control room modernization of a significant system in the plant, especially one such as a safety system that requires a license amendment with the regulator and compliance with NUREG-0711 (O'Hara et al., 2012), will generally benefit by availing itself of all phases of the GONUKE process. Conventionally, Validation will provide the most directly conclusive results, while Verification and Epistemiation provide supplemental evidence on the successful execution of the system design. As noted, ISV should not be the sole form of evaluation used.

After the GONUKE process has been followed once, later design changes may not require all phases of evaluation. A small change to the system may benefit from revisiting the Summative evaluation, while a large-scale change may require design iterations aligned to revisiting Formative evaluation.

A simplified version of the GONUKE process (Boring, Joe et al., 2014) relevant to most design evaluation processes is found in Table 4. This simplified version of GONUKE suffices to establish a good design when the rigors of formal regulatory review are not required, such as when a non-safety-critical system is modernized in the control room or only graded approach is feasible due to budget or time constraints. The simplified version of GONUKE ensures that the critical steps in evaluation are considered as a design is finalized. This simplified version may also prove the most generalizable rendition of the approach for non-nuclear applications.

Table 4. Simplified Usability Evaluation Types and Phases for Non-Safety-Critical Systems.

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

(This page intentionally left blank)

5. NEW DIRECTIONS FOR VERIFICATION AND VALIDATION IN NUCLEAR POWER PLANTS

5.1 Introduction

The Safety Case (see Chapter 3) provides an overarching framework for structuring evaluation in terms of requirements, arguments, and evidence that help ensure safety assessment and prediction are not built solely only on summative evaluation (i.e., ISV). GONUKE (see Chapter 4) presents a process containing twelve evaluation stages defined along the dimensions of evaluation phase and type (see Table 3) for highlighting the human factors methods applicable at different points of the modernization project life cycle. While both the Safety Case and GONUKE deserve further research, they represent methodological developments in research and practice that can guide the industry toward a view that V&V should be concerned with safety outcomes over the entire licensing period rather than a singular type of testing activity. At the same time, substantial methodological development in other areas remains necessary to support the nuclear industry to adopt such V&V approaches. This chapter presents research recommendations that would enable the nuclear industry to move beyond the conventional V&V activities.

5.2 Improving Evidence Collection

Given the guidance of the Safety Case and GONUKE for structuring the evidence and evaluation process into twelve stages, the next pertinent research area is identification and development of human performance metrics for the corresponding evaluation stages. That is, research needs to provide technical guidance on collecting specific evidence that can be structured according to the Safety Case and GONUKE.

Research that deserves immediate attention is (1) assigning available human performance metrics and (2) guiding application of those metrics appropriately for individual stages. Human performance metrics or measurement systems have been developed for the nuclear industry (see Table 5). Further, classic metrics such as response time to alarms and time to actions sometimes provide meaningful indicators of human performance, although time-based metrics are sensitive only in scenarios with highly time-critical events (Skraaning Jr., 2003). However, most human performance metrics or measurements are developed or employed mainly for ISV, namely testing operators in high-fidelity simulators; thus, direct application of these human performance metrics and measures may be unfeasible or incompatible for some evaluation stages, even if they are otherwise well-established in the nuclear industry. Other human factors assessment techniques or metrics may be suitable for some evaluation stages but lack supporting research in the nuclear domain. In essence, research on methods is necessary to apply the available measurement methods to collect the evidence corresponding to individual evaluation stages to build a safety case.

Table 5. Measurement Research in the Nuclear Domain.

Methodological Research/Techniques	Type	Description
Operator Performance Assessment System (OPAS; Skraaning Jr., 1998, 2003; Skraaning Jr. et al., 2007)	Task performance metrics	Expert rating of performance items based on observation
Self-rated task performance	Task performance metrics	Self-rating of task performance (Demas, Lau, & Elks, 2015)
Halden Complexity Measure (Braarud, 2000; Braarud & Brendryen, 2001)	Workload metrics	Questionnaire on scenario challenges and workload specific for the nuclear industry
NASA TLX (Hart & Staveland, 1988)	Workload metrics	Questionnaire on workload for all industries and accepted in the nuclear domain
Situation Awareness Control Room Inventory (SACRI; Hogg, Follesø, Torralba, & Volden, 1994; Hogg, Follesø, Volden, & Torralba, 1995)	Situation Awareness (SA) metrics	A query-based measure of SA for the nuclear industry based on the Situation Awareness Global Assessment Technique (SAGAT; Endsley, 1995)
The Process Overview Measure (Lau et al., 2014; Lau et al., 2011a, 2011b)		Evolutionary advancement of SACRI
Human Performance Evaluation Support System (HUPESS; Ha, Seong, Lee, & Hong, 2007; Ha & Seong, 2009)	Human performance measurement system	A framework systematically describing importance and integration of various measures

The particular challenge of adopting measures originally developed for ISV to pre and post-ISV applications is simplification of the various measurement methods and metrics for deployment to a range of scenarios. As mentioned in Chapter 2, increasing methodological complexity tends to impede deployment of measures, especially in applied practice. For early evaluation, general rather than precise indication of control room operation is sufficient. Simplification may be achieved by developing a standard suite of scenarios that suitably couple with various validated metrics. Research should also identify the set of control room features that can be validated individually, leaving summative evaluation for assessing critical interactions between various operational elements of the control room.

In addition to well-established methods, new measurements that are non-intrusive to collect and easy to analyze require development. In particular, physiological measures such as eye-tracking and breathing rate may hold promise if coupled with appropriate standard scenarios. Eye-tracking research in process control is already showing promise (e.g., Gao, Wang, Song, Li, & Dong, 2013; Ikuma, Harvey, Taylor, & Handal, 2014) but further research is necessary to turn eye-tracking measurements into easy to analyze and interpret metrics for control room evaluation representative of industrial settings (see Demas et al., in press).

5.3 Improving Prediction

The admission of all evidence that is predictive of safety along the entire project cycle should improve confidence in the V&V decision outcome. However, V&V must account for the fact that individual pieces of evidence are not equal in merits for safety prediction. For early V&V activities to be appreciated in the design and licensing process, the merits of different evidences at various evaluation stages need to be clearly specified. In other words, the decision consequence for evidence of different forms (e.g., SA vs workload metrics) and of different evaluation stages must be explicit in order to guide control room design and safety prediction.

To help assess merits of different evidence types (e.g., metrics at a particular stage), future research could examine measurement errors and performance predictability of different evidence. In addition to the varying conceptual importance for different performance categories, both qualitative and quantitative data likely contain varying levels of measurement errors that affect data analysis and conclusions. Both process experts as raters and operators as study participants involved in data collection and/or analysis methods produce varying level of errors depending on the circumstances (e.g., scenarios and measures). For example, measurement errors for a specific measure within a single process expert are probably not constant. Process experts might be more accurate and reliable in estimating or judging operator performance for within-design-basis than beyond-design-basis scenarios. The rationale of this statement can be deduced from gross over- and under- estimation of event probabilities for extremely common and rare events (e.g., Wickens, Hollands, Parasuraman, & Banbury, 2012). Further, most people consistently underestimate probabilities for frequencies in the mid range. Discovering the tendencies of process experts and/or operators in performance interpretation and design feedback is equally applicable to qualitative as well as quantitative data analysis. In summary, improving knowledge on measurement (and interpretation) errors associated with process experts for individual measures, conditions, and technology types can improve the confidence in any performance and safety prediction. As noted by Boring (2004), there is a tendency to use humans as measurement devices in psychological research, but there is little effort to calibrate those humans—a unique shortcoming that hinders effective use of data collected by process experts.

If the assumption is valid that measurement errors vary with (some) scenario types, then the methods of data collection and analysis deserve strategic selection. For instance, small measurement errors for particular type of scenarios afford few data points to achieve the necessary confidence (though not necessarily statistical significance). In contrast, large measurement errors for some scenario types deserve increased sampling and caution. Further, operator performance for handling familiar events may orient towards quantitative methods while performance for handling unanticipated events may orient towards qualitative methods.[6] In other words, knowledge of measurement errors can inform the methods of collecting and the amount of quantitative and qualitative evidence for different types of operating circumstances or other aspects of safety, especially when data points are fewer than necessary to provide conclusive (conventional) inferential statistics.

Empirical research on measurement error variations with respect to process experts (e.g., Lau et al., 2014), performance categories (e.g., Lau, Jamieson, & Skraaning Jr, 2012), and operating conditions can provide correction factors to operator and system performance assessment. This approach is equally applicable to early-stage evaluation in terms of calibrating operator feedback on design. The literature includes many examples in which operator preference does not lead to

[6] This idea should be subjected to debate.

the optimal design. Unfortunately, research on measurement errors in the nuclear domain is too limited to provide any serious guidance on how to select measurement methods or how to correct for the errors. The stellar safety record in the nuclear industry can be largely attributed to the conservative safety culture and deep domain knowledge of the nuclear industry when assessing operator performance and system safety. In other words, the high confidence in the safety of nuclear power operation is a result of the conservative culture rather than our scientific accuracy of safety assessment. Further, safety culture and domain knowledge are likely to influence V&V measurements and decision validity in the future.

Confidence in V&V outcomes or licensing decisions thus rests on understanding the measurement errors or variability in the data collection and analysis methods. That is, our knowledge of methods (and the domain) can be and probably is being applied to moderate the confidence levels in individual performance results (cf., probability estimates in human reliability analysis). Thus, research that can provide an empirical foundation for estimating the errors or variability of various performance testing conditions and measures would improve confidence in V&V conclusions and decisions. The improved confidence stems from an empirical basis for moderating or correcting the individual performance estimates (or claims) for predicting operational safety and thus making valid licensing decisions.

A feasible research strategy to study measurement variability and predictability under different testing conditions would be invaluable for building an empirical basis that assesses merits of V&V performance results from small-sample-size testing[7]. The proposed strategy for acquiring knowledge on measurement variability is to collect (or simply record) qualitative and/or qualitative performance data during simulator training required for licensed operators several weeks per year (e.g., five weeks in the US). In fact, this strategy is already proposed by a recent human reliability analysis research program called Scenario, Authoring, Characterization, and Debriefing Application (SACADA; Chang et al., 2014)[8]. This strategy is feasible because much of the work for the performance testing is already being (and must be) done as part of recurring required operator training in control room simulators. The key missing element in simulation training sessions at nuclear power plants appears to be formalizing the measures and recording the data. If the utilities would implement well-established or test novel human performance data collection methods and share the data in their simulation training sessions, sufficient data can be feasibly collected to study human performance measurement errors, particularly in expert judgment variability across a multitude of test factors (such as scenario types). Further, the data bank of collected data can provide a reference performance level that puts a particular data point

[7] The notion of measurement errors and predictive validity may have a connotation employing quantitative metrics in summative evaluation. However, the concept is applicable for qualitative feedback on design during early V&V. For example, operators may be asked to provide qualitative feedback on the performance impact of a technology and the predictability of safety regarding this feedback may be assessed in terms of the performance in a training session when the particular technology fails.

[8] This approach is similar to the Strategic Highway Research Program 2 (SHRP2) naturalistic driving study that instrumented 3,000 vehicles to observe and collect data on ordinary people about their driving behaviors. The data bank supports calculating odds ratios of a particular behavior leading to crashes (relatively rare events).

collected in V&V performance testing into context. That is, the collected data would have a corresponding performance and variance level for comparison.

Collecting and recording human performance data or operator inputs during simulator training sessions offers benefits that could increase confidence in drawing performance conclusions in V&V. Collecting human performance data provides reference performance and variability levels for many common scenarios of performance tests, which individually cannot provide a large enough sample size suitable for generalization. Collecting operator inputs provide a reference consensus regarding the importance and relevance of various control room features and technology. Such reference performance or feedback offers an empirical basis for qualitatively and quantitatively evaluating the V&V results, thereby improving the confidence (or consequential validity) in the V&V decision. In addition, a priori hypotheses can be formulated with respect to the reference performance levels if necessary.

In relation to the first benefit, the available data provide indication on when expert judgment and corresponding interpretation become less reliable, yielding weak performance prediction. Isolating less reliable performance results can improve the V&V conclusion and decision validity. By identifying the conditions with poor reliability, research and V&V efforts can be allocated accordingly. For instance, V&V efforts may focus testing on beyond-design-basis scenarios once comparable performance and variability levels become apparent for within-design-basis scenarios. Alternatively, V&V can mainly target new failure modes that do not exist in the original plants.

Standardization in collecting and recording human performance data in already formal simulator training sessions provides benchmark performance for plants undergoing modernization. The benchmarks enable other data collection and analysis methods, including single-case experimental designs (SCED) that rely on visual inspection of performance data (Rizvi and Ferraioli, 2012). The medical field employs SCED frequently to study drug intervention efficacy, but the method requires careful measurements of baseline behaviors. If utilities undergoing modernization are actively collecting performance data during simulator training sessions, SCED can be effectively employed for V&V. Some research on this method in the nuclear domain is likely warranted. In any case, collecting human performance data during simulator training can provide the additional data to facilitate test designs that can improve confidence in drawing performance conclusions.

Another benefit mainly relevant for modernization projects is the exposure of process experts to using or implementing current human performance data collection methods. As mentioned, the quality of human performance measurements depends heavily on working with process experts. Increased exposure to using and working with human performance measures reduces time demands to produce quality data when performing V&V. Quality data improve confidence in performance conclusions (and statistical power for quantitative analysis).

Given that only limited data can be feasibly collected from a specific performance test, performance conclusions in V&V often cannot practically be drawn according to criteria designed for scientific pursuit (mostly in academia). In the nuclear industry, V&V results for predicting future performance become heavily dependent on the domain experience of the process experts and other professionals to fill data voids. Thus, one strategy to increasing the confidence in performance conclusions would be to learn about the test conditions in which experts make reliable or unreliable performance estimation and prediction. Further, reference performance levels would be invaluable to put individual performance data points into context. Studying measurement variability and providing reference performance levels can feasibly be accomplished by collecting qualitative or quantitative performance data during simulator training sessions necessary for licensing individual operators. That is, simulator training sessions can help generate a human performance data bank that could help interpret individual performance data points collected from V&V performance testing. This data bank would help estimate the level of confidence in a specific V&V finding and thus the specific performance prediction.

6. CONCLUSIONS

Every system that is designed for use by operators should be validated and verified to ensure it meets safety objectives and is usable. Despite the importance of V&V, it is an activity that is often relegated to the late stages of the design process in safety-critical domains like control room modernization. In this report, we've made the case for using V&V across the design lifecycle, especially at early stages when V&V can positively shape the design of the system. The advantages of early-stage evaluation include not only the ability to improve the design but also to ensure operator buy-in and to avoid potential reworks of the system that might be necessary when issues are first discovered late in the design and development process.

Despite these advantages, the true value of early-stage V&V should also be understood in terms of building the case for the safety of the system. Evidence for the safety of the system should not be limited just to ISV. Adoption of iterative design and evaluation demonstrates a solid HFE process and should serve to establish confidence by the utility and the regulator that error traps have been eliminated as the design has matured. Early-stage V&V coupled with late-stage ISV forms a comprehensive picture of the safety of the system. Following the stages of evaluation outlined in GONUKE will ensure that safety concerns have been identified and mitigated during the design lifecycle. Although there is no requirement for V&V outside ISV in NUREG-0711, the process outlined in this report fully supports regulatory goals of the new design.

We close this report with two final considerations about V&V as it pertains to control room modernization.[9]

- Modern DCSs used for control room modernization are much less static than their analog predecessors. They can be fine-tuned over time for better performance, better display presentation, or better alarm management. The HSIs run on standard operating systems and will need continuous maintenance. Ongoing changes to the digital architecture and feature set suggest the need for additional V&V, even after the ISV is completed. The scope of gradual HSI changes to control systems may not warrant a large-scale design activity that steps through all stages of NUREG-0711. The use of early-stage V&V methods may translate into a sustainable approach for ensuring safety of systems as they are gradually upgraded. The model of a large evaluation for a large change in the control room may not hold when the changes become more nuanced. Without a suite of methods to assess these changes quickly and cost effectively, there is the risk that small changes may not take advantage of V&V. Small-scale, discount evaluation such as suggested by ALARA may be the key to ongoing evaluations that mirror the natural evolution of digital HSIs.

- It should be remembered that V&V is a confirmatory approach. It is only intended to show that operators can use the system for prescribed conditions. As such, HFE, operations, and engineering need to be diligent in casting a wide net in selecting scenarios. Still, it is never possible to anticipate all possible scenarios. The role of ISV is to test the integration of the tested new system against the other systems with which it interacts. In this manner, system dependencies and common cause failures can be identified. Early-stage V&V presents a

[9] These ideas originated in a conversation with Dr. Roger Lew of the University of Idaho.

> different type of confirmation. Early-stage evaluation tends to be more informal and open-ended, exploring operator first interactions with the system across unscripted activities. In many cases, early-stage V&V precedes operating procedures, thereby necessitating a degree of discovery by the operators. This discovery may actually be seen as a type of stress test of the system as operators familiarize themselves with the system interface and its strengths and weaknesses. The opportunity to gather performance data on first and unconventional use scenarios can actually instill confidence in the robustness of the system. Insights from early-stage evaluation are crucial in establishing a pattern of interaction that can be extrapolated to novel and even unanticipated scenarios. The safety of the system is not just proved through carefully considered scenarios; it is ultimately demonstrated through the system's resilience across diverse uses including those that are unforeseen. Early-stage V&V represents an ideal test case for the system outside normal operations.

Control room modernization has only begun to realize the benefits of V&V. V&V is often considered within a narrowly defined function to support ISV. Expanding the application of V&V promises to create an integrated design process that can become the backbone of plant safety assurance. V&V should become a continuous process as plants modernize, providing graded levels of evaluation suitable to support both small and large upgrades at various stages of design and deployment.

7. REFERENCES

Boring, R.L. (2004). Cognition and Psychological Scaling: Model, Method, and Application of Constrained Scaling. PhD dissertation, Institute of Cognitive Science, Carleton University, Ottawa, Canada.

Boring, R.L. (In press). Envy in V&V: New directions for verification and validation in nuclear power plants. Submitted for the Human Factors and Ergonomics 59th Annual Meeting.

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

Boring, R.L., Agarwal, V., Joe, J.C., & 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.L., & Bye, A. (2008). Bridging human factors and human reliability analysis. Proceedings of the 52nd Annual Meeting of the Human Factors and Ergonomics Society, 733-737.

Boring, R.L. & Joe, J. C. (2014). Baseline Human Factors and Ergonomics in Support of Control Room Modernization at Nuclear Power Plants, INL/EXT-14-33223, Idaho National Laboratory, Idaho Falls, 2014.

Boring, R., Joe, J., Ulrich, T., & Lew, R. (2015). Operator Performance Metrics for Control Room Modernization: A Practical Guide for Early Design Evaluation, INL/EXT-14-31511, Rev. 1. Idaho Falls: Idaho National Laboratory.

Boring, R.L., Joe, J.C., Ulrich, T.A., & Lew, R.T. (2014). Early-stage design and evaluation for nuclear power plant control room upgrades. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 58, 1909-1913.

Boring, R.L., Jung, W., Lau, N., & Skraaning, G. (2015). How to run a control room simulator research study. Proceedings of the American Nuclear Society Nuclear Plant Instrumentation, Control, and Human Machine Interface Technology (ANS NPIC & HMIT) Conference, pp. 1560-1568.

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

Boring, R.L., Tran, T.Q., Gertman, D.I., & Ragsdale, A.S. (2006). A human reliability based usability evaluation method for safety-critical software, ANS NPIC-HMIT. pp. 1275-1279.

Boring, R.L., Ulrich, T.A., Joe, J.C., & Lew, R.T. (In press). Guideline for operational nuclear usability and knowledge elicitations (GONUKE). Proceedings of the 6th International Conference on Applied Human Factors and Ergonomics (AHFE).

Braarud, P. Ø. (2000). Subjective task complexity in the control room (HWR-621), Halden, Norway: OECD Halden Reactor Project.

Braarud, P. Ø., & Brendryen, H. (2001). Task demand, task management, and teamwork (HWR-657), Halden, Norway: OECD Halden Reactor Project.

Chang, J. Y., Bley, D., Criscione, L., Kirwan, B., Mosleh, A., Madary, T., . . . Zoulis, A. (2014). The SACADA database for human reliability and human performance. Reliability Engineering & System Safety, 125(0), 117-133.

Demas, M., Lau, N., & Boring, R. (In press). Eye Tracking Applications in Formative Evaluation of Control Room Modernization: A Center for Advanced Engineering and Research Technical Report Prepared for Idaho National Laboratory. Idaho Falls: Idaho National Laboratory.

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

Diaper, D. & Stanton N. (2004). The Handbook of Task Analysis for Human-Computer Interaction, Lawrence Erlbaum Associates, Mahwah, NJ.

Electric Power Research Institute (2005). 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, EPRI TR-1010042. Palo Alto, CA: EPRI.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. Human Factors, 37(1), 65-84.

Fuld, R.B. (1997) Verification and validation: what’s the difference? Ergonomics in Design, 5(3), pp. 28–33.

Gao, Q., Wang, Y., Song, F., Li, Z., & Dong, X. (2013). Mental workload measurement for emergency operating procedures in digital nuclear power plants. Ergonomics, 56(7), 1070-1085.

Ha, J. S., & Seong, P. H. (2009). HUPESS: Human Performance Evaluation Support System. In P. H. Seong (Ed.), Reliability and Risk Issues in Large Scale Safety-critical Digital Control Systems (pp. 197-229): Springer London.

Ha, J. S., Seong, P.-H., Lee, M. S., & Hong, J. H. (2007). Development of Human Performance Measures for Human Factors Validation in the Advanced MCR of APR-1400. Nuclear Science, IEEE Transactions on, 54(6), 2687-2700.

Haddon-Cave, C. (2009). The Nimrod Review: An Independent Review into the Broader Issues Surrounding the Loss of the RAF Nimrod MR2 Aircraft XV230 in Afghanistan in 2006. London: Office of Public Sector Information.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N.

Meshkati (Eds.), Human Mental Workload (pp. 139-183). Amsterdam, The Netherlands: Elsevier Science Publisher.

Hogg, D. N., Follesø, K., Torralba, B., & Volden, F. S. (1994). Measurement of the operator's situation awareness for use within process control research: Four methodological studies (HWR-377),Halden, Norway: OECD Halden Reactor Project.

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

Hollifield, B.R., Oliver, D., Nimmo, I., & Nabibi, E. (2008). The High Performance HMI Handbook. A Comprehensive Guide to Designing, Implementing, and Maintaining Effective HMIs for Industrial Plant Operations. Houston: PAS.

Hopkins, A. (2013). The Cost-Benefit Hurdle for Safety Case Regulation. Washington, DC: U.S. Chemical Safety Board.

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

Ikuma, L. H., Harvey, C., Taylor, C. F., & Handal, C. (2014). A guide for assessing control room operator performance using speed and accuracy, perceived workload, situation awareness, and eye tracking. Journal of Loss Prevention in the Process Industries, 32, 454-465.

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

Joe, J.C., Boring, R.L., & Persensky, J.J. (2012). Commercial utility perspectives on nuclear power plant control room modernization. 8th International Topical Meeting on Nuclear Power Plant Instrumentation, Control, and Human-Machine Interface Technologies (NPIC&HMIT), 2039-2046.

Kane, D. (2003). Finding a place for discount usability engineering in agile development: Throwing down the gauntlet. Proceedings of the IEEE Agile Development Conference, pp. 1-7.

Kelly, T. (2008). Are ‘safety cases’ working? Safety Critical Systems Club Newsletter, 17(2), 31-33.

Kelly, T.P. (1998). Arguing Safety—A Systematic Approach to Managing Safety Cases. PhD Dissertation. York, UK: University of York Department of Computer Science.

Lau, N., Jamieson, G. A., & Skraaning Jr, G. (2012). Inter-rater reliability of expert-based performance measures. Paper presented at the Proceedings of the 8th American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation & Control and Human-Machine Interface Technologies (NPIC & HMIT), San Diego, CA, USA.

Lau, N., Jamieson, G. A., & Skraaning Jr., G. (2014). Inter-rater reliability of query/probe-based techniques for measuring situation awareness. Ergonomics, 57(7), 959-972.

Lau, N., Skraaning Jr, G., Eitrheim, M. H. R., Karlsson, T., Nihlwing, C., & Jamieson, G. A. (2011a). The Process Overview Measure: Methodological developments to enhance inter-rater reliability (HWR-971), Halden, Norway: OECD Halden Reactor Project.

Lau, N., Skraaning Jr, G., Eitrheim, M. H. R., Karlsson, T., Nihlwing, C., & Jamieson, G. A. (2011b). Situation awareness in monitoring nuclear power plants: The Process Overview concept and measure (HWR-954), Halden, Norway: OECD Halden Reactor Project.

Lindgaard, G., Dillon, R., Trbovich, P., White, R., Fernandes, G., Lundahl, S., & Pinnameneni, A. (2005). User needs analysis and requirements: Theory and practice, Interacting with Computers 18. pp. 47-70.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. Educational Measurement: Issues and Practice, 14(4), 5-8.

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

Murphy, K.R. (2009). Validity, Validation and Values. Academy of Management Annals, 3(1), 421-461.

Nielsen, J. (1995, January). Applying discount usability engineering. IEEE Software, pp. 98-100.

O'Hara, J.M., Brown, W.S., Lewis, P.M., & Persensky, J.J. (2002). Human-System Interface Design Review Guidelines, NUREG-0700, Rev. 2. Washington, DC: U.S. Nuclear Regulatory Commission.

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

O'Hara, J.M., Higgins, J.C., Persensky, J.J., Lewis, P.M., & Bongarra, J.P. (2004). Human Factors Engineering Program Review Model, NUREG-0711, Rev. 2. Washington, DC: U.S. Nuclear Regulatory Commission.

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

Office of Nuclear Regulation (2013). The Purpose, Scope, and Content of Safety Cases, Health and Safety Executive, London.

Persensky, J., Boring, R., Le Blanc, K., Hugo, J., Gertman, D., Shaver, E., Braun, C., Oxstrand, J., & Richards, R. (2010). Alarm System Research Plan: Milestone Deliverable in Support of the U.S. Department of Energy's Light Water Reactor Sustainability Project, INL/EXT-10-19888. Idaho Falls, ID: Idaho National Laboratory.

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

Rizvi, S. L., & Ferraioli, S. J. (2012). Single-case experimental designs. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological. (pp. 589-611). Washington, D.C., USA: American Psychological Association.

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

Russell, B. (1961). The Basic Writings of Bertrand Russell. London: George Allen & Unwin Ltd.

Scriven, M. (1967). The methodology of evaluation, in: R.E. Stake (Ed.), Curriculum Evaluation, Rand McNally, Chicago.

Skraaning Jr., G. (1998). The Operator Performance Assessment System (OPAS) (HWR-538), Halden, Norway: OECD Halden Reactor Project.

Skraaning Jr., G. (2003). Experimental control versus realism: Methodological solutions for simulator studies in complex operating environments. (HPR-361), Halden, Norway: OECD Halden Reactor Project.

Skraaning Jr., G., Lau, N., Welch, R., Nihlwing, C., Andresen, G., Brevig, L. H., . . . Kwok, J. (2007). The ecological interface design experiment (2005) (HWR-833), Halden, Norway: OECD Halden Reactor Project.

Strobhar, D.A. (2014). Human Factors in Process Plant Operation. New York: Momentum Press.

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

U.S. Food and Drug Administration. (2010). Infusion Pumps Total Product Life Cycle: Guidance for Industry and FDA Staff, OMB 0910-0766. Washington, DC: Center for Devices and Radiological Health, U.S. Department of Health and Human Services, U.S. Food and Drug Administration.

UK Ministry of Defence. (2015). Safety Management Requirements for Defence Systems, Part 1: Requirements and Guidance, Defence Standard 00-56 Part 1, Issue 6. London: UK Ministry of Defence.

UK Office for Nuclear Regulation. (2013). The Purpose, Scope, and Content of Safety Cases, NS-TAST-GD-051 Rev. 3. Bootle: UK Office for Nuclear Regulation, Health and Safety Executive.

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.A. & Boring, R.L. (2013). Example user centered design process for a digital control system in a nuclear power plant, Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting. pp. 1727-1731.

Ulrich, T.A., Boring, R.L., & Lew, R. (2015). Control board digital interface input devices—Touchscreen, trackpad, or mouse? Proceedings of 2015 Resilience Week, 168-173.

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

Wickens, C. D., Hollands, J. G., Parasuraman, R., & Banbury, S. (2012). Engineering psychology and human performance (4th ed.). Upper Saddle River, NJ: Pearsons.